NASALITY IN AUTOMATIC SPEAKER VERIFICATION

Edmund Joseph Rooney

Thesis submitted for the degree of Ph.D.
University of Edinburgh
1990

ABSTRACT

This thesis examines the suitability of nasal resonance patterns as a means of authenticating speakers' identities in an automatic speaker verification system.

The inadequacy of traditional methods of ascertaining identity in commerce and industry – the possession of keys or PIN numbers, for example – has prompted researchers to look at attributes which are inseparable from the person who possesses them ("*biometric*" features: that is, features which are part of a person's physical make-up, or aspects of their performance of a task). The use of speech in this application has received much attention, despite its inherent variability. Much of the research uses whole-word templates (text-dependent) or long-term statistical measures (text-independent), but a third approach – segmental analysis – has also proved useful, because it concentrates on features of speech which are known to be highly speaker-dependent.

The nasal cavities in particular are known to vary considerably from speaker to speaker, and to be relatively fixed in their size and shape. The acoustic analysis of nasality is complicated by the manner of its production, however, which introduces *anti-resonances* or transfer function *zeros* into the spectrum. This renders the most popular analysis technique, Linear Predictive Coding, inherently inaccurate, since it assumes a vocal tract transfer function which has all poles (resonances) and *no* zeros. In this thesis, the potential of nasality is re-examined using a relatively new but established technique, *cep-*

stral decomposition, which allows accurate estimation of both pole and zero frequencies. The efficacy of this technique is demonstrated on both synthetic speech and nasal stops, and a modification is introduced to reduce the detrimental effects of overestimation of the all-zero model order.

A review of acoustic, anatomical and phonetic aspects of nasality suggests that while nasality does not offer an invariant acoustic marker of identity (the nasal tract proving extremely variable and its contribution to the spectrum depending extensively on the rest of the vocal tract), it still offers the most favourable phonetic environment for the purposes of speaker verification. The velar nasal stop is chosen for study, since its spectrum shows the greatest dependence on unalterable nasal tract characteristics and the greatest resistance to changes elsewhere in the vocal tract (e.g. lingual coarticulation).

Acoustic variability in the velar nasal is explored on a hand-segmented database of thirty speakers, with recordings spanning eight weeks. The effects of sex, vowel context (coarticulation), speaker identity and intra-speaker variability on resonance and anti-resonance frequencies obtained by cepstral decomposition are considered. A method of *peak profile warping* is proposed to overcome the problem of misalignment of spectral peak and dip frequencies across tokens.

The viability of using the velar nasal spectrum for speaker verification is explored using resonance and anti-resonance frequencies and the pole-zero spectrum itself. Context-dependent reference formation, variance-based feature weighting and Canonical Analysis are all found to improve the results.

Verification equal error rates of 12.8% (for 15 males) and 13.8% (for 14 females) are achieved using references formed from four enrolment sessions and single test utterances taken from separate test sessions over four weeks. Speakers are shown to vary considerably in their consistency. The results of an adaptation experiment show that a relatively crude form of adaptation is also beneficial.

Some observations are made on the possible success of practised impostors at imitating nasal resonance patterns, and on the need for more detailed study of the effects of physiological change.

ACKNOWLEDGMENTS

I acknowledge gratefully the many contributions of others to the completion of this thesis. In particular, my thanks are due to:-

- my supervisors, Prof. John Laver and Prof. Mervyn Jack, who have guided me in my work and given freely of their time and wisdom, and who as Chairman and Director of the Centre for Speech Technology Research have done so much to stimulate us all;

- Mr. Alan Kemp, for his advice and interest;

- my colleagues on the De La Rue Speech Technology Project, 1985-1988 – Pam Rodriguez, Andrew Sutherland, George Fletcher and Shona Anderson – with special thanks to Shona, who gave many hours of her time to digitize the database, and to Andrew, who provided the error rate software and much useful criticism, and helped to smooth the way;

- my colleagues in the Centre for Speech Technology Research and the Department of Linguistics, and their spouses, who provided recordings for the speaker database collected as part of that project;

- the ever-patient Computing Officers at CSTR – Andie Ness, Tim Bradshaw and Alex Zbyslaw – and Bob Anstruther, our Communications Technician;

- Prof. B. Yegnanarayana, for his helpful and encouraging comments on Chapter Four of this thesis;

- Dr. Clive Summerfield for the implementations of the pole-zero decomposition technique and the peak-picking algorithm used in this thesis;

- my present colleagues on the Integrated Speech Technology Demonstrator Project at CSTR – Alan Wrench, Fergus McInnes, Steve Hiller, Alan Crowe and Harold Blackburn – for a great deal of advice and encouragement, and much tolerance;

- Ellen Bard for her advice on statistical matters on countless occasions;

- the De La Rue Company, for their financial support of the project from which this thesis developed;

- the British Academy, for financial support early in my postgraduate career;

- my family, especially my father, who have supported me and never given up hope;

- my friends, especially Rosemarie, who has never given up asking;

- and Cathy, who has waited a long time for this and always provided love and encouragement, and to whom, with Aidan, this thesis is dedicated.

DECLARATION

This thesis was composed by me and reports original work of my own execution, part of which formed my contribution to a research group investigating Automatic Speaker Verification on behalf of the De La Rue Company.

ABBREVIATIONS

ARMA	Autoregressive Moving Average
cm	centimetre
cps	cycles per second
dB	decibel
DFT	Discrete Fourier Transform
DPS	Derivative of Phase Spectrum
EER	Equal Error Rate
FA	False Acceptance (Rate)
FFT	Fast Fourier Transform
FR	False Rejection (Rate)
HMM	Hidden Markov Model
Hz	Hertz
IPA	International Phonetic Association
kHz	kilohertz
LAR	Log Area Ratio
LP	Linear Prediction
LPC	Linear Predictive Coding
MAE	Minimum Average Error
MER	Minimum Error Rate
ms	milliseconds
NDPS	Negative Derivative of Phase Spectrum
PARCOR	Partial Correlation
PCM	Pulse Code Modulation
PDA	Pitch Determination Algorithm
rms	root mean square
RP	Received Pronunciation
s.d.	standard deviation
SRM	Speaker Recognition by Machine
SRS	Speaker Recognition by Spectrogram
SRL	Speaker Recognition by Listening
VQ	Vector Quantization

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

CHAPTER ONE

INTRODUCTION

1.1. Background and aims

This thesis examines the suitability of nasal resonance patterns as a means of authenticating speakers' identities in an automatic speaker verification system.

The need to ascertain people's identity is particularly felt in the world of business and industry, where large amounts of money, sensitive information or military or industrial secrets or machinery are at risk. Traditional methods of ensuring their security – the use of keys, keycards and passwords – are becoming inadequate, since they rely on the possession of objects to give authority. Physical objects such as keys and keycards are easily lost or stolen, while codes or passwords which must be remembered can be forgotten, or divulged to others.

This inadequacy has prompted researchers to look at attributes which are inseparable from the person who possesses them. In particular, there is a great deal of interest in so-called "biometric" features – that is, features which are part of a person's physical make-up, or aspects of their performance of a physical task.

Fingerprinting is the most striking example of a biometric technique; up to now its primary use has been in forensic science, but systems are now available which allow a person's fingerprint to be used instead of a signature in support of credit transactions, for example. A more restricted technique is the use of the retinal pattern – an image of the network of blood vessels visible on the retina of the eye. Both methods have the advantages that they exploit a person's physical features directly, and that as far as is known these features are unique, unchanging and unalterable.

Aspects of a person's performance of a physical task are less directly related to their physical characteristics, but are still considered to have a biometric basis. The use of personal handwriting, particularly signatures, is already commonplace, and much research has gone into the construction of automatic methods to remove the need for subjective (and patently fallible) human judgement.

All these methods have disadvantages, though, which restrict their possible applications. Fingerprinting and the use of retinal patterns both require special optical scanners, and intimate physical contact with the person. In the case of retinal scanning, this may be distressing or embarrassing. Signature verification may be ideally suited to some applications, such as credit authorisation in shops, but is impractical elsewhere, especially where physical access is required.

Verification from a person's speech, though more loosely a biometric technique, has been the subject of extensive research. It has great flexibility, and is

suited to a wide range of applications. One of its major advantages is the ease with which speech samples can be obtained: all that is required at the point of contact with the speaker is a microphone – something which is familiar to most people and cheap to instal. Verification can also be done remotely, over telephone lines or radio links; and with the widespread use of "Entryphone" systems in blocks of flats, the idea of establishing a person's identity by their voice alone is becoming as familiar as the use of signatures.

The major problem with voice-based verification is the inherent *variability* of speech. Speech is not a fixed physical object, but a complex form of behaviour which reflects the speaker's anatomy and physiology only indirectly, and which by its very nature is constantly changing, since without change no meaning could be conveyed. The vocal tract is in *constant* motion during speech, even during apparently "steady-state" articulations. In addition, the speech apparatus itself and the speaker's control of it are subject to change, from day to day (as with the effects of health, fatigue and psychological states) and in the long term (as with the effects of disease or ageing).

Various approaches have been tried to lessen the effects of this variability. One is to look for features which reflect long-term "settings" of the vocal tract, or some form of statistical description which is independent of the content of the speech and therefore of the changing sequence of segments. Another is to use the patterns of change themselves, by matching "contours" or "templates" of whole utterances. A third, which has received much less attention, is to focus on features of speech which show the greatest dependence on the physical

make-up of the speaker, and to study them over relatively short intervals, during which the vocal tract is reasonably stable.

The use of segmental nasality is an example of the third approach. The nasal cavities are known to vary considerably from speaker to speaker, and to be relatively fixed in their size and shape. The acoustic analysis of nasality is complicated by the manner of its production, however, which introduces *anti-resonances* or transfer function *zeros* into the spectrum. This renders the most popular analysis technique in verification studies, Linear Predictive Coding, inherently inaccurate, since this technique assumes a vocal tract transfer function which has all poles (resonances) and *no* zeros. Estimates of anti-resonance frequencies are thus quite vague, while spectral peaks are perturbed from their true positions. The relatively low energy in nasal segments (resulting from high energy losses in the nasal tract) also makes analysis difficult. The study of nasality has therefore been neglected in recent years, while techniques in speaker verification and basic analysis methods have been improving.

In this thesis, then, the potential of nasality is re-examined, and a relatively new technique for characterising it – pole-zero decomposition – is applied to the problem. This technique, developed by Yegnanarayana (1981), uses Fourier-based cepstral analysis to estimate separate all-pole and all-zero frequency responses for the vocal tract over a given stretch of speech. This allows accurate estimation of spectral peaks and dips independently of each other, and produces a spectrum which is a more accurate model of the vocal tract response.

1.2. Structure of the thesis

Chapter Two provides a general review of the field of speaker verification, and highlights the relative neglect of nasality. The nature of the verification task, the structure of a basic verification system, and the main methods of decision-making are described. The choice of suitable parameters for speaker verification is discussed, and the range of approaches to the verification problem – fixed-text or free-text, text-independent, text-dependent and semi-text-dependent – is described. The main part of this chapter is devoted to consideration of the different parameter sets available – phonation parameters, spectral parameters and timing parameters – and describes some of the results obtained in the literature. Finally, the use of nasality in speaker verification is reviewed.

A full discussion of nasality is presented in **Chapter Three**, paying particular attention to anatomical and physiological variability and how it is reflected in acoustic variability. The production of nasality is reviewed, and it is demonstrated that not all manifestations of what is called nasality depend on the nasal cavities, and that the conditions for nasal resonance are complex. The role of nasality in language is considered, to show that it is almost universal as a segmental feature. Nasal stops are seen to provide the best exponents of nasality for the purposes of verification.

The anatomical variability of the nasal cavities is then reviewed; it is seen that, while they are the most fixed of all the vocal tract cavities in their overall dimensions and shape, and show enormous between-speaker differences, their

cross-sectional area and volume are capable of rapid physiological change. The acoustic theory of nasality is then described, and the acoustic characteristics of nasal stops and nasalized vowels outlined. The acoustic properties of the vocal tract cavities, including the nasal cavities, are considered, along with their behaviour in normal speech, particularly with reference to lingual coarticulation during nasal stops. While the velar nasal stop emerges as the best candidate, showing greatest resistance to coarticulation and the maximum dependence of all the stops (except the uvular stop) on the nasal tract, it is seen that there is very little data on its acoustic structure and variability.

Chapters Four, Five and Six attempt to provide some data on the acoustic variability of the velar nasal using an analysis technique which is suited to nasality: the pole-zero decomposition method of Yegnanarayana (1981).

Chapter Four provides a review of some of the analysis methods proposed for accurate modelling of signals in which spectral zeros as well as poles occur, highlighting the inadequacy of Linear Prediction Analysis. The reasons for the choice of the pole-zero decomposition method are presented, and a pilot experiment using a synthetic speech signal is used to show the greater accuracy of this technique over Linear Prediction.

Its suitability is explored using real speech tokens – nasal stops – in **Chapter Five**. Data obtained on the pole and zero (or peak and dip) frequencies are compared against published values, with encouraging results. The need to reduce the modelling order for the all-zero part of the model is shown.

Chapter Six uses the pole-zero decomposition method to obtain data on spectral peak and dip frequencies for 15 male and 15 female speakers, using velar nasal tokens recorded over a period of between eight and ten weeks. The effects of the sex difference, vowel context, speaker identity and intra-speaker variability are considered. A method of *peak profile warping* is proposed to overcome the problem of misalignment of peak and dip frequencies prior to statistical analysis.

Chapter Seven examines the viability of using the velar nasal spectrum for speaker verification. The basis for evaluating the success of a system is discussed first, and two suitable measures of distance are described. The peak-dip features obtained in Chapter Six are compared with the combined pole-zero spectrum, which proves to be superior. The performance of the pole-zero spectrum is then examined under various conditions. Context-dependent reference formation and comparison, variance-based feature weighting and Canonical Analysis are all found to improve the results. The distribution of errors over the speakers is examined, and speakers are shown to vary considerably in their consistency. Finally, the results of an adaptation experiment are reported, showing that a relatively crude form of adaptation is also beneficial.

Chapter Eight, the final chapter, gives a brief summary of the research findings, highlights areas in which further work might be useful and improvements obtained, and discusses three major areas not dealt with in this thesis: the use of automatic segmentation for locating the nasal stops, the effects of physiological variation (head-colds, infections and so on), and the possible suc-

cess of imitation of nasal resonance patterns by trained impostors.

The transcription system used throughout this thesis is the Machine Readable Phonetic Alphabet (MRPA), developed at the Centre for Speech Technology Research, University of Edinburgh; the principal symbols and their IPA equivalents are reproduced in Appendix A on p.344.

CHAPTER TWO

AUTOMATIC SPEAKER VERIFICATION – A REVIEW

CHAPTER TWO

AUTOMATIC SPEAKER VERIFICATION – A REVIEW

2.1. Introduction

This chapter contains a general review of the field of speaker verification, to provide a background for my own work. Automatic speaker verification is defined with reference to the wider area of speaker recognition, and a typology of approaches to the problem of verification is outlined. Systems and studies described in the literature are then examined in the light of this typology. All studies reviewed here have in common the fact that they exploit features of low-level acoustic detail, rather than higher level linguistic features. It will be clear from this review that nasality has been neglected, especially in recent years.

Several reviews of this field have already appeared over the last twenty years. These include Atal (1976), Corsi (1981), Doddington (1985), Furui (1986), Hecker (1971), Jesorsky (1978), Nolan (1983), O'Shaughnessy (1986), Rosenberg (1976) and Sutherland and Jack (1988). This chapter therefore does not attempt to provide an exhaustive treatment.

2.2. Definitions

Automatic speaker verification

Automatic speaker verification is part of the general field of speaker recognition – the study of how speakers can be recognised from the characteristics of their voice. Speaker recognition has been defined as

> any decision-making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance

(Atal 1976: 460). *Automatic* speaker recognition is the use of machines – typically computers – to "recognize" speakers from samples of their speech: these samples are processed to give a small set of features based on some acoustic characteristics of the speaker's voice (for example, their average pitch), and these features are compared with stored sets of features derived from one or more known speakers to give a decision on the identity of the unknown speaker.

Automatic speaker *verification* is the name given to a particular type of recognition in which a speaker first makes an explicit identity claim and then provides the computer with a voice sample to allow that claim to be checked. One important application of automatic speaker verification is in access control systems, in which entry to premises is controlled by computer: a person wishing to gain access to a secure room or building identifies himself to the computer by inserting a magnetic card with a personal code number on it, or by entering a personal identification number (PIN) on a key-pad; he is then asked to speak a short phrase or a few isolated words; this speech sample is processed

by the computer, and if the features derived are sufficiently similar to the features of the person he claims to be, the computer unlocks the door. Similar systems can be used to control access to money, for example from cash dispensers, or to information and services such as computer files. The use of the voice in this way has some important advantages over alternative methods relying solely on keys or key-cards (which can be lost or stolen), or on memorised code numbers (which can easily be forgotten). It is also ideal for telephone transactions such as credit card sales, in which a person's signature – the normal means of verifying their identity – cannot be checked.

Speaker recognition

The range of processes covered by the label "speaker recognition" is very wide indeed. Nolan (1983) has provided a useful basis for classifying these processes. His primary distinction is between *naive* speaker recognition and *technical* speaker recognition. Naive speaker recognition is the term given to the natural human ability to recognize speakers. It has also been termed "auditory speaker recognition" (Brown 1980) and "speaker recognition by listening" (Hecker 1971). Nolan's term is to be preferred, though, because it prevents possible confusion with *technical* methods involving auditory analysis.

Technical speaker recognition brings together a range of skills and techniques which do not rely on this natural ability, but which instead make use of specialized training or analysis methods, applied in a conscious, methodical fashion. The best-known and most controversial example of such a technique is the recognition of speakers from "Voiceprints" or sound spectrograms – visual

representations, produced by spectral analysis, of the sound energy present in an utterance (Tosi 1979).

This division between naive and technical speaker recognition is illustrated in Figure 2.1. Within what Nolan calls technical speaker recognition there are several methods which can be graded by the degree of human involvement in the recognition process. At one end of the scale are the fully *automatic* methods – those in which all aspects of recognition (extraction of relevant

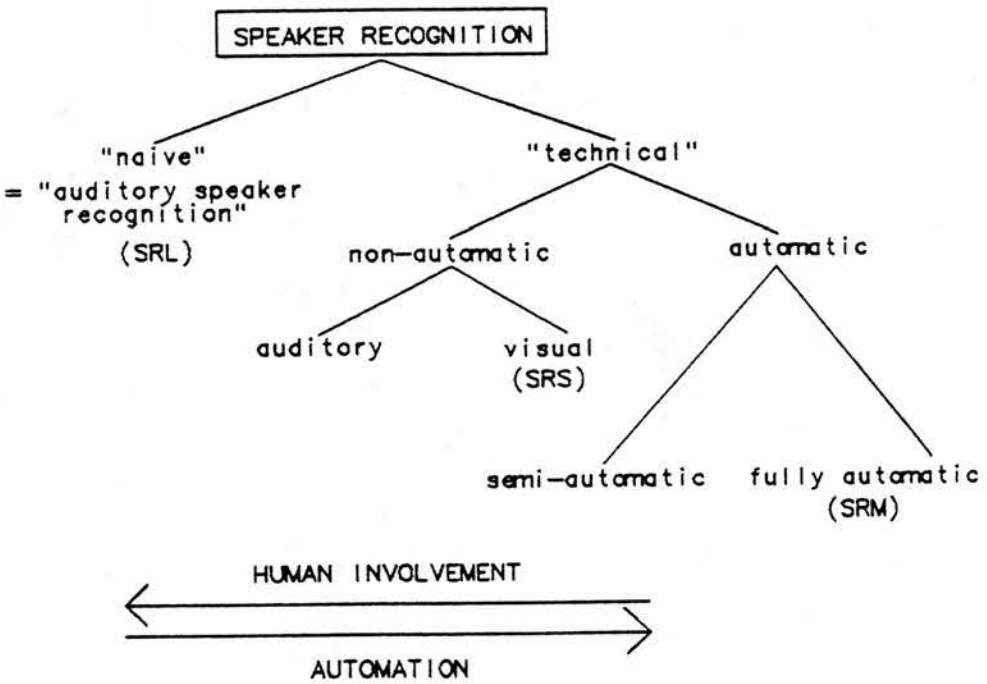

Figure 2.1 Categories of speaker recognition

features from speech and a recognition decision) are performed by machine. Then there are the so-called *semi-automatic* methods, in which the speech is prepared for computer analysis by human operators, who use their own skills and judgement to decide which parts of an utterance are most relevant. Together these correspond with Hecker's (1971) category of "Speaker Recognition by Machine" or SRM: in both cases, the essential parts of the recognition process – those of comparing voices and coming to a decision – are performed by machine, according to a set of explicit rules. At the other end of the scale are what may be termed (rather vacuously) *non-automatic* methods, comprising visual recognition (from speech spectrograms) and skilled auditory analysis by phoneticians. Only the former is considered by Hecker (his category of "Speaker Recognition from Spectrograms" or SRS). Skilled auditory recognition is a separate category from naive auditory recognition because it involves the systematic application of analysis techniques acquired during phonetic training. Spectrogram analysis involves the same mechanised analysis process as many automatic methods, but the recognition process is fully under the control of a human expert.

Speaker verification and speaker identification

Technical speaker recognition as defined above subsumes two distinct tasks: speaker *verification* and speaker *identification*.

Speaker verification – the topic of this thesis – involves checking a speaker's explicit claim to a particular identity. The decision required of the

system is one of *discrimination* – deciding whether speaker A (the claimant) and speaker B (the named person) are the same. There are four possible outcomes to such a decision: correct acceptance of a genuine claimant, correct rejection of an "impostor", false rejection of a genuine claimant and false acceptance of an "impostor". The last two outcomes are the error conditions known as "Type 1" (FR) and "Type 2" (FA) respectively (Doddington 1976: 398).

Speaker identification, on the other hand, involves assigning an utterance of unknown origin to one speaker in a group of known speakers. The decision required is one of *categorisation* (or *classification* – there is no generally accepted term). This decision usually has just two outcomes: correct assignment of the unknown utterance to the right speaker, and incorrect assignment of the utterance to another speaker.*

2.3. Basic properties and structure of automatic speaker verification systems

2.3.1. Modes of operation

A working speaker verification system such as the Texas Instruments (Doddington 1985) or Bell Laboratories system (Rosenberg and Sambur 1975), operates in two basic modes: (a) enrolment and (b) verification.

(a) During *enrolment* reference materials for a set of authorised speakers are stored in computer files. These reference materials generally comprise, for

* This represents the case known as "closed-set" identification, in which it is assumed that the unknown speaker is a member of the group of known speakers. A third outcome is possible where there is the chance that the utterance came from a speaker not represented in the group of known speakers, that is, in the case of "open-set" identification.

each authorised speaker, a set of measurements on one or more analysis parameters (a *feature vector*), derived from speech samples provided by that speaker. Once the system has representative references for all authorised speakers, it can be used to verify unknown speakers.

(b) Each time a speaker presents himself for verification, he identifies himself to the system by means of a personal code (entered on a keypad, for instance, or read from the magnetic strip on a swipe-card). This code tells the system which set of reference measurements will be needed. A speech sample is taken from the unknown speaker (perhaps in response to a visual prompt), and is analysed to give a similar set of parameter measurements. The two sets of measurements – the reference and the "bid", as the unknown speaker's contribution is commonly termed – are compared by the system, which has to decide whether they are similar enough to have come from the same speaker.

2.3.2. System structure

The elements of any verification system are *pre-processing*, *feature extraction* and *feature comparison* (Jesorsky 1978). The place of these operations can be seen in Figure 2.2, which shows a typical system, having both enrolment and verification functions. Pre-processing and feature extraction are the operations which are common to both enrolment and recognition; feature comparison is, as is clear from the description above, only required in the verification

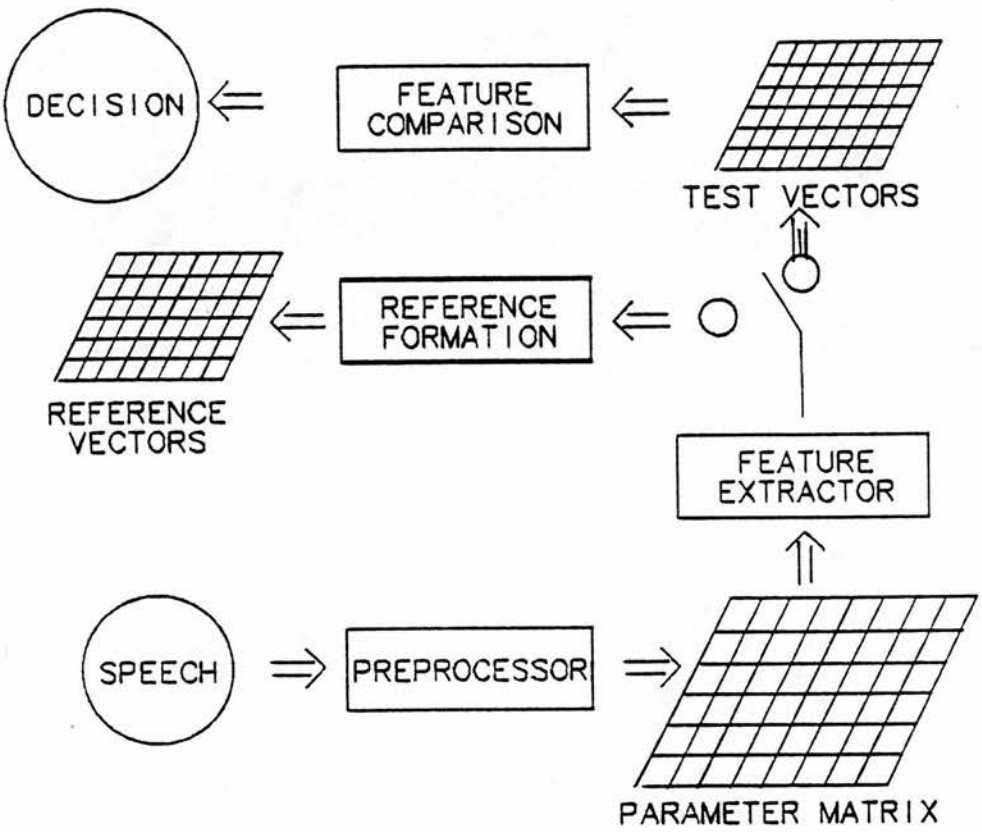

Figure 2.2 The main features of an automatic speaker verification system

mode.*

In preprocessing, the acoustic speech signal is converted to a parametric representation by some form of speech analysis. Many analysis techniques exist for this "parameterisation". All have in common the fact that their outputs vary much more slowly than the original speech waveform: they can therefore be sampled relatively infrequently without significant loss of informa-

* It may be used in sophisticated reference formation methods, though (e.g. Fakotakis et al. 1987)

tion, giving a discrete representation from which speaker-characterising features can be extracted at the next stage.

At the stage of feature extraction, representative measurements are made on the discrete parametric representation to give a *feature vector*:

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]$$

For example, the fundamental frequency parameter (showing the changes of voice pitch over time) may be represented by a feature vector containing just its mean (or other measure of average value) and its variance; or a set of values may be extracted at particular points during the utterance (such as during stressed vowels, or at the beginning and end of the utterance only), and used to form the feature vector on their own. A common approach, however, is to use the discrete parametric representation directly, without a separate feature extraction stage.

The reference data for a speaker typically comprise some representative *average* of a set of feature vectors provided during enrolment. Some information must also be provided on the typical distribution of a speaker's feature vectors relative to the feature vectors of other speakers, to allow the classifier to decide whether a test vector belongs to that speaker or not.

The final stage, feature *comparison* or classification, involves the comparison of the feature vector derived from the unknown speaker's bid with the reference feature vector stored in the system's memory. Comparison is frequently done by measuring the "distance" between the reference vector and the bid vector, as if the two were points in multidimensional space. A threshold is

applied, and if the resulting distance is above this threshold the speaker's claim is rejected.

2.3.3. Determining the acceptance/rejection threshold

Errors in verification occur either when a genuine bid from an authorised speaker produces a distance which exceeds the acceptance threshold (Type 1 error, or False Rejection); or when a bid from another speaker produces a distance which falls below this threshold (Type 2 error, or False Acceptance). The threshold is determined during training or enrolment from the distributions of intra-speaker distances (distances resulting from comparisons of a speaker's own tokens with their reference) and inter-speaker distances (those resulting from between-speaker comparisons). Figure 2.3 (a) shows the frequency distributions of intra-speaker distances and inter-speaker distances for a hypothetical population. Generally, intra-speaker distances will be smaller than inter-speaker distances. Figure 2.3 (b) shows the same data in cumulative form, with the intra-speaker distance distribution inverted. As the distance threshold along the abscissa increases, the percentage of intra-speaker tokens showing distances above this threshold decreases, as does the probability $P(FR)$ of falsely rejecting one of these genuine tokens. Increasing the distance threshold even further removes entirely the risk of false rejection, but now some inter-speaker comparisons yield distances below the threshold, and the risk of false acceptance $P(FA)$ begins to rise. Thus the setting of the acceptance threshold determines the balance between the two types of error.

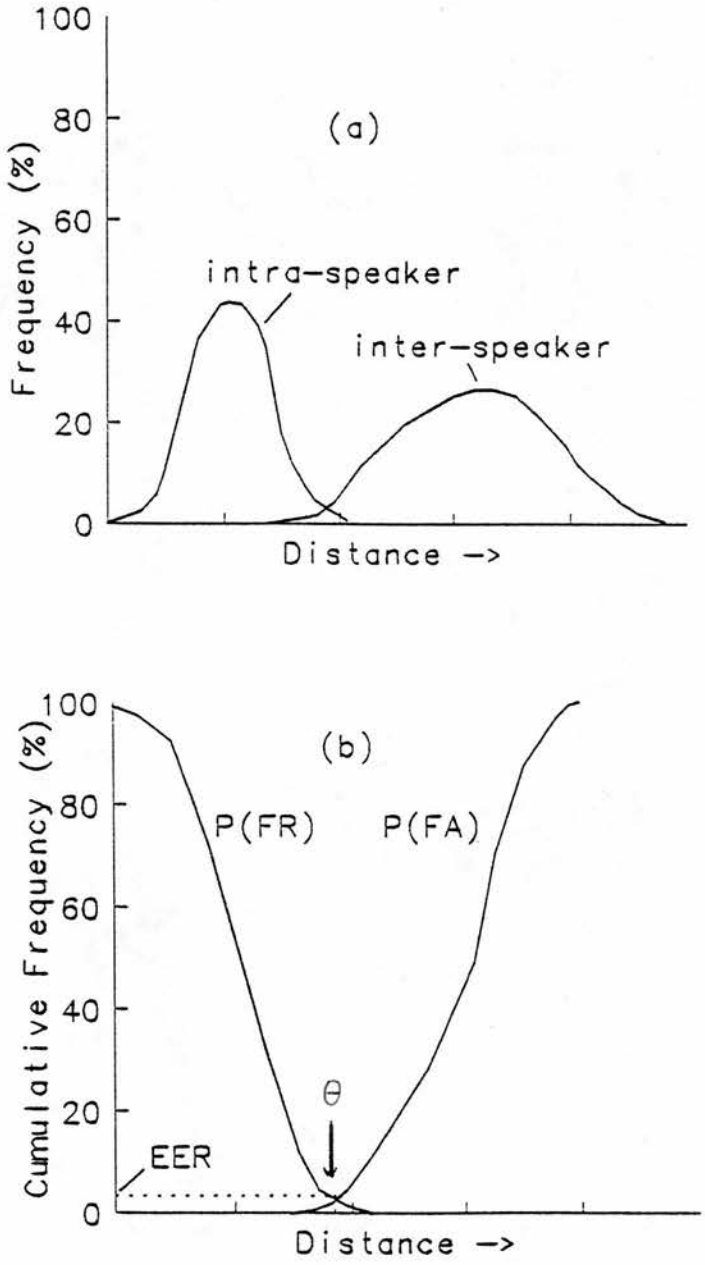

Figure 2.3 Intra-speaker and inter-speaker distance distributions (schematic), showing the derivation of the a-posteriori EER threshold

The acceptance threshold must be set at a level which gives the most tolerable combination of False Rejection and False Acceptance – a combination

whose balance depends on practical matters such as the degree of security required and the costs of denying entry to authorised personnel. One compromise frequently adopted is to set the threshold at the level which gives an equal probability (on the basis of the training data distance distributions) of False Acceptance and False Rejection. This point is found where the two cumulative frequency (or probability) curves cross, and is known as the equal error rate threshold (Θ on Figure 2.3 (b)). The corresponding error rate (the *Equal Error Rate* or EER) can be read from the ordinate.

2.3.4. Measuring a system's performance

If the training data are truly representative of the verification population, then the application of the threshold Θ to distances calculated from an independent data set should give a similar result – that is, approximately equal levels of Type 1 and Type 2 errors, at around the level obtained on the training data. Very few studies test their system on independent data with the threshold preset in this way, however; examples are Das and Mohn (1971) and Furui 1981b. Instead, the Equal Error Rate determined from the training data (the *a-posteriori* EER, as it is known, because it is derived from a threshold imposed on the distances once their distribution is known) is itself used as an indicator of the future performance of the system (e.g. Doddington 1971, Markel and Davis 1979, Rosenberg and Soong 1986).

A less common alternative to the EER point is the point at which the sum of the two types of error (or their average) reaches a minimum (the Minimum Error Rate, or MER). This is less useful as an indicator of performance,

however, since it is influenced heavily by the shape of the distributions of inter-speaker and intra-speaker distances.

Some studies (e.g. Fakotakis and Kokkinakis 1985) have indicated performance simply by the number of errors occurring, expressed as a percentage of the total number of verification attempts, a method commonly used in speaker identification (e.g. Hunt et al. 1977). This method gives a misleading picture if applied to verification systems, however, since the errors can be of two types.

2.3.5. Feature Selection

The above description of the operation of a speaker verification system presupposes that a suitable set of parameters and features has been decided upon. The selection of such a set of features is an essential part of the design phase. It should not be confused with feature extraction. A typical approach to the problem is to choose a set of possible features, bearing in mind the application the system will serve, and to evaluate this set on a group of speakers. The aim of doing this is to choose a subset of features which will discriminate speakers as efficiently as possible.

There are many methods of selecting features in use, and many of the speaker verification studies reviewed in section 2.7 are devoted to evaluating different feature subsets. One method is to choose features which show a large F-ratio in an analysis of variance: that is, those with a high degree of between-speaker variance relative to their within-speaker variance (Wolf 1972). The ideal method, as Atal (1976) points out, is to choose the feature subset which minimises the empirically-determined error rate on a test set of

speakers. However, the number of feature combinations is often very large, and calculating error rates in this way is very time-consuming and wasteful of data.

Some systems (e.g. Sambur 1976, Furui 1981b) incorporate a feature selection stage into their enrolment of speakers, so that the efficiency of the feature subset is reassessed each time a new speaker is registered. Normally, however, feature selection is limited to the design phase of the development of a system.

2.4. The choice of parameters for speaker verification

A number of factors govern the choice of parameters used for speaker verification and recognition in general. Nolan (1983), revising an earlier formulation by Wolf (1972), summarizes them as follows:

- high between-speaker variability
- low within-speaker variability
- resistance to attempted disguise or mimicry
- availability in speech
- robustness in transmission
- measurability

As Nolan observes, however, most studies in speaker recognition are simply evaluations of an existing set of parameters, and very few studies consciously apply such criteria to select their parameters.

Between-speaker and within-speaker variability

The variability of parameters from speaker to speaker and within the performance of a single speaker is the principal criterion for selecting parameters in speaker verification: generally, features are chosen which show the largest

differences between speakers, but which are stable for any given speaker. However, the *sources* of variability in speech are still poorly understood, and have received relatively little attention beyond a rather crude categorization into *organic* and *learnt* differences (Garvin and Ladefoged 1963, Stevens 1972, Wolf 1972, Atal 1976). Organic differences are those which stem from the physical constraints imposed by a speaker's vocal tract and lungs; learnt differences are those which reflect "acquired" speaking habits, whether idiosyncratic or accent-related. It has been suggested that features which reflect physical differences will be the most reliable recognition clues of all, since physical characteristics cannot be altered at will, are relatively constant over time, and vary greatly from speaker to speaker. Their effects are found mainly at the lowest levels of speech production, influencing the acoustic realisation of speech sound segments and the coordination and timing of the four component processes of sound production: airstream generation, phonation, articulation and nasality (Ladefoged 1971). Features which have been suggested include spectral features such as formant frequencies of vowels, fricatives and nasal stops, which are heavily dependent on the dimensions and structure of the vocal tract (e.g. Glenn and Kleiner 1968, Wolf 1972, Sambur 1975) and fundamental frequency distributions, dependent on the size and mobility of the vocal folds (e.g. Clarke and Becker 1969).

Nolan points out, however, that this organic-learnt distinction is of limited use, since the plasticity of the vocal tract means that there are few if any cases of an organic feature leaving an invariant acoustic mark; at most, the

speaker's organic apparatus determines the *limits* of his production. For example,

> There may be a physiologically determined maximum and minimum to a given speaker's fundamental frequency range; and his preferred range may in some sense be the optimal one given his particular larynx; but he nevertheless has at his disposal a variety of other fundamental frequency ranges within the absolute physiological limits

(1983: 27-28).

It is clear from Nolan's (1983) review of speaker variation that there is enormous potential for variation at *all* levels of speech production, in both the segmental and suprasegmental strands. Not all of these sources of variation are relevant in the fairly constrained circumstances of automatic speaker verification: higher level choices of syntax and lexicon are outwith the speaker's control, for example, because the text is usually determined in advance. Some factors must still be taken into account, though. At the level of *communicative intent*, for example, factors such as the speaker's mood, their confidence in using the system, or the need to project an appropriate "image" of themselves in the presence of other people, may affect the choice of suprasegmental features such as pitch range, and segmental features such as phonetic realisations (the choice between a glottal stop and a voiceless alveolar stop, for example). The level of background noise may influence the speaker's voice level, affecting both pitch control and spectral slope.

Resistance to mimicry

Resistance to *disguise* is not relevant to verification, since speakers seeking access will not normally try to conceal their own identity other than by

attempting to impersonate someone else already known to the system. Resistance to *mimicry*, however, is a critical property, but one which is seldom explored. Studies generally simulate attempts at impersonation using data from "casual" impostors – speakers who provide tokens of the verification utterances without attempting to imitate any of the authorised speakers (e.g. Rosenberg 1976, Doddington 1985). This approach gives a rather optimistic picture, as the few studies which have used "real" impostors, attempting to impersonate authorised speakers, have demonstrated. Lummis and Rosenberg (1972) gave a group of selected professional mimics intensive training (with auditory feedback and information on the distances achieved) on 8 genuine speakers from Doddington's (1971) 40-speaker database. The best utterances from the best four mimics were processed by Doddington's system, and yielded a False Acceptance rate of 27%, compared with a rate of 1.2% for casual impostors. Rosenberg and Sambur (1975) found a smaller increase in the False Acceptance rate, from 1% to 4%, when mimics were used. Finally, in a study by Hair and Rekeita (1972), a single professional mimic was allowed to familiarise himself with 6 speakers from the 40-speaker database used in Hair and Rekieta (1972a); the mimic then produced imitations of each speaker immediately after they had uttered a selected word. Spectral analysis of individual segments showed that the mimic achieved increased similarity for some speakers on some phonemes, but in a verification test using a combination of five phonemes (not specified) the mimic was rejected each time.

Some resistance to mimicry can be achieved by choosing speech parameters which depend on features of speech production that cannot be altered at will, such as physiological attributes (Rosenberg 1976), or which are not perceptually salient and are therefore likely to be overlooked by impostors (Nolan 1983). Lummis and Rosenberg (1972) found that formant information gave some protection against trained impostors, though an earlier study (Doddington 1971) had shown that with casual impostors it was no better than the parameters of fundamental frequency and gain.

Rosenberg (1976) points out that experiments with impostors are difficult to interpret, since the definition of a skilled mimic varies according to the requirements of each system. This makes the choice of mimics very difficult, since it is hard to know in advance which speakers will be able to give convincing-enough imitations to fool the system. He observes, however, that mimics who depend on caricature are unlikely to be successful, "since most effective systems will not tolerate exaggeration of speaker characteristics" (1976: 480).

Availability

The situation in which verification is performed gives a certain amount of control over the *availability* of the speech features of interest, since the text to be uttered is chosen by the system, and can be made to include the desired features no matter how uncommon they are in everyday speech. However, they should still be able to be extracted from a relatively short utterance, and this rules out many long-term statistical parameters, which require upwards of 10

seconds of speech to reach stability. If segmental features are chosen, care must be taken that they occur in all the dialects to which the system might be exposed.

Robustness in transmission

Nolan points out that parameters which are removed or distorted by the process of recording or transmission (down telephone lines, for example) should be avoided. In the case of telephone speech this is imperative, since the nature of the distortion imposed on the signal changes with the telephone connection and the equipment used (Furui 1977, Moye 1979, Bogner 1981). The effects of telephone distortions on verification performance have been studied extensively. McGonegal, Rosenberg and Rabiner (1979) found that prosodic parameters (fundamental frequency and gain) were less severely affected than spectral parameters. Several studies have attempted to compensate for the spectral changes introduced, by extracting some measure of the channel characteristics themselves (Ichikawa et al. 1978); in the case of cepstral analysis, removal of the long term average of each cepstral coefficient usually has this effect (Furui 1981a). Hunt (1983), in an identification study, found that spectral *peak* features such as formant frequencies had a greater resistance to noise and non-linear spectral distortions than spectral parameters such as cepstral coefficients. In verification for access control, though, the transmission characteristics are under control and generally constant, and known distortions can therefore be compensated for.

Measurability

The measurement and extraction of the parameters chosen should obviously not be too difficult. In verification applications, the time allowed for measurement is usually rather short, and there is a preference for parameters which can be derived in real time by dedicated hardware (such as filter-banks) or fast digital analysis such as Linear Prediction. This is one reason for the lack of studies using Fourier analysis, which takes rather longer to perform (Furui 1981a). A large number of studies save time by using the analysis parameters directly, without extracting features. Systems using a segmental approach are at a disadvantage, since they need an extra stage at which the appropriate segments are located. Errors at this stage render feature extraction and comparison impossible, and may contribute significantly to the error performance of the system, as with Das and co-workers' study (1971), in which there was no decision in 10% of cases because of problems such as segmentation errors.

2.5. The uses of speech material for verification

The main determinant of the nature of the speech materials used is the application to which the system will be put. In general, designers of verification systems for highly-controlled environments can choose speech materials to suit their design; whereas designers of recognition systems for use in uncontrolled environments frequently have to choose their design to suit the speech materials. The validity of system trials depends on the use of appropriate speech materials for the application.

2.5.1. Free-text versus fixed-text

One major consideration is whether the text on which the recognition system must operate can be prescribed by the operator (*fixed-text*) or is free to vary (*free-text*). Verification systems for access control (e.g. Doddington 1976, 1985; Bielby et al. 1987, and Feix and De George 1985) generally expect fixed-text utterances: the use of a fixed text helps to reduce the amount of processing required, guarantees the presence of the features of interest, and can reduce intra-speaker variability, as speakers become familiar with the utterances they are required to give. This can only be done because speakers are cooperative. Other types of recognition system must be designed to cope with more variable speech input. For instance, recognition systems designed for undercover surveillance (for military or criminal intelligence, for example) must operate on whatever the target speakers care to say, and no assumptions can be made as to the textual content of utterances, nor even as to their length: these are true free-text systems (e.g. Li and Wrench 1983, Wolf et al. 1983, Krasner et al. 1984).

2.5.2. Types of utterance

Fixed-text systems, especially in speaker verification, typically use short utterances such as single sentences (e.g. Lummis 1973, McGonegal et al. 1979), phrases (e.g. Naik and Doddington 1986, 1987), or isolated words (e.g. Bielby et al. 1987). Utterances must be kept short in access-control applications, to minimise delays. The use of a short fixed text has the disadvantage of making imitation or pre-recording of an authorised speaker much easier, however, so a

common approach is to use a restricted set of utterances from which a prompt can be chosen at random during the verification attempt (Doddington 1976, 1985). The text is still fixed in the sense that the verification system knows what to expect from the speaker (assuming the speaker follows the prompt faithfully).

Free-text systems, by definition, cannot know in advance what the content of the speech will be, and must be designed to cope with a variety of speech styles and utterance durations. Realistic testing of such systems is difficult. Relatively few studies use truly unconstrained speech: notable examples are that by Krasner and co-workers (1984), in which speech samples range from 0.5 to 5 seconds in length, and the study by Markel and Davis (1979), using speech from interviews up to 13 minutes long. Commonly, fairly long recordings of read speech are used in system trials (e.g. Hollien and Majewski 1977, Schwartz et al. 1982, Johnson et al. 1984), but this does not offer the same variability as that found in spontaneous speech. The study by Hunt (1983) demonstrates this (and the effects of recording conditions): an identification accuracy rate of 89% achieved on a database of good quality read speech fell to 66% on unrestricted telephone conversations.

2.5.3. Content

All studies reviewed in this chapter exploit acoustically-measurable features of the speech signal: thus the cognitive (and even semantic) content of the speech materials is largely irrelevant, except so far as it affects speakers' attitudes to what is being said (if it is emotive, humorous, embarrassing or

stressful, for example), or encourages unusual or idiosyncratic interpretations of what sort of emphasis and intonation are appropriate. Similarly, the language and accent used for a study will have relevance only to the availability of the particular features chosen. Most studies use English (British or North American), but examples of other languages studied are German (Höfker 1977), Telugu (Pal and Majumder 1977), Mandarin (Chen and Lin 1987), Japanese (Furui 1981b), Italian (Frederico et al. 1987), and Greek (Fakotakis et al. 1987).

The actual content of the texts presented to speakers varies enormously, and frequently owes more to practical considerations than to considerations of phonetic theory. Thus digit names ("*zero*" to "*nine*") are very commonly used in verification studies (e.g. Buck et al. 1985, Naik and Doddington 1986, 1987), partly because of a widespread interest in digit-sequence recognition (for computer data entry, for telephone dialling, and in financial transactions); Rosenberg and Shipley (1983) actually base their verification decision on the prior operation of an isolated word recognition system. Bielby and co-workers (1987) included words such as "*add*", "*change*", "*erase*" and "*stop*", presumably with computer text-processing in mind. Ease of acoustic analysis is an important consideration, too: Feix and De George (1985), for example, chose words beginning with plosives and ending with non-plosives to facilitate word-boundary location; and sentences and phrases are often chosen to contain only phonemically voiced segments (e.g. Furui 1981a, Lummis 1973), to simplify the task of formant trackers or pitch determination algorithms and to maximise the

number of voiced analysis frames available to the feature extraction stage. Some studies select only oral phonemes, (e.g. Lummis 1973) since nasal segments can cause problems for some analysis techniques: Rosenberg and Sambur (1975), for example, using Linear Prediction, found that an all-oral sentence "*We were away a year ago*" gave fewer verification errors than a sentence containing nasal segments, "*I know when my lawyer is due*".In some cases (e.g. Atal 1972, Doddington 1976, 1985) *nonsense* phrases have been used, but this is not wise when fundamental frequency variations are being studied (as in Atal 1972), since speakers will tend to vary in their interpretation of what stress pattern and intonation are appropriate (Nolan 1983).

Studies which are interested in individual segments, such as vowels (e.g. Frederico et al. 1987, Fakotakis and Kokkinakis 1985), and nasal stops (Das and Mohn 1971, Kashyap 1976) need to include these segments in their materials, but normally pay little regard to phonetic context or syllable position.

The speech to be read is almost always presented visually, but some workers (e.g. Sambur 1975, Doddington 1976, 1985) have used voice prompting in an attempt to stabilise speakers' pronunciations by giving them a model to aim at, though no comparisons are available which might tell us whether this is effective.

2.6. Text-dependent and text-independent systems

Speaker verification systems are often referred to as either "text-dependent" or "text-independent". This distinction – which is subtly different from that between "fixed-text" and "free-text" systems – relates to the method

of feature extraction and comparison.

There are two general approaches to feature extraction and comparison: *text-dependent* operation and *text-independent* operation. Both attempt to deal with the fact that the acoustic parameters of speech reflect not only the desired specific characteristics of the speaker, but also the phonemic *content* of the utterance (that is, the identities of the speech sounds themselves) and the influence of all the higher-level (syntactic and semantic) information the utterance contains.

2.6.1. Text-dependent operation

In text-dependent operation, no attempt is made to separate the characteristics of the speaker from those of the utterance; instead, the phonemic content of the utterance is held constant during both training and verification, and analysis parameters or features from test utterances are compared directly with stored patterns for the corresponding reference utterance. For this method to work, it is necessary to align the test and reference utterances so that the comparison is made between parameter values at the same points in each utterance. This is especially important in the case of spectral parameters, which vary greatly from segment to segment: comparing the spectrum of a voiceless fricative in the test utterance against that of a vowel in the reference utterance is not valid.

This alignment is achieved in different ways. One method involves linearly compressing or stretching one utterance until it is the same length as the other, or at least has the same number of analysis points. Atal (1974), for

example, divided every utterance of the sentence "*May we all learn a yellow lion roar*" into 40 frames of equal duration, no matter what its length, so that while the frame length varied from utterance to utterance, the number of elements in each analysis was the same, allowing a direct comparison. This does not guarantee a perfect alignment between corresponding points, however, since the effects of lengthening or compressing an utterance in speech are *not* linear: vowels, for example, typically lengthen more than consonants. An alternative method is to compress or expand different parts of the utterance as needed, until the best match is obtained. This method is known as *dynamic time warping* (Sakoe and Chiba 1978) and has been used extensively in speaker verification studies based on parameter *contours*. Lummis (1973), for example, compared the contour of the fundamental frequency parameter extracted from the sentence "*We were away a year ago*" with the corresponding contour produced by each speaker during enrolment by matching the two gain contours. In Doddington's system (1976, 1985) words are represented by just six spectral vectors spanning 100 milliseconds, taken from the middle of the word (presumably centred around the point of maximum intensity); time alignment is achieved simply by choosing from the middle of the test word the six vectors which give the lowest squared Euclidean distance from the reference segment. This method is similar to the segmental analysis discussed below.

2.6.2. Text-independent operation

The second major approach – *text-independent* operation – attempts to derive from the parametric representation of each utterance a set of features

which characterise the speaker independently of the content of the utterances used and without the need for alignment of corresponding features. Thus the content of the utterances used for enrolment and of those used during recognition need not be the same; nor do utterances have to be warped to identical lengths. The features used tend to be *statistical*, describing the distribution of values found for a given parameter over the duration of the utterance. Such features tend to be less dependent on the phonemic content of the utterance than the parameter contours themselves, but for some parameters such as fundamental frequency (used by Lummis (1973), Sambur (1975), Furui (1981b) and Chen and Lin (1987)) quite a large amount of speech is needed (in excess of 60 seconds: Nolan 1983) before speakers' distributions stabilise.

The approach chosen depends partly on the type of application the system is designed for. Text-independent methods may be used in any application, but *must* be used where speakers are not expected to be cooperative, or where fixed text cannot be expected. However, they generally give poorer results than text-dependent methods under the same circumstances (e.g. Rosenberg and Soong 1986), and are sometimes not suitable for access-control applications since they require rather long utterances. Text-dependent methods, though slightly superior (since some of the variability has been removed), can be used only when the speaker can be expected to be cooperative and a fixed text is possible; they give better results than text-independent methods on short text, and are therefore better suited to access-control.

2.6.3. Semi-text-dependence

There are three variations on the text-dependent, contour-based approach to verification which reduce the need for alignment and free the system from strict dependence on a fixed text. The first of these involves isolating particular phonetic segments or acoustic events in an utterance for comparison with stored feature values – the *segmental* approach. The second and third variations, *Vector Quantisation* and *Hidden Markov Modelling* do away with this requirement for isolating segments by performing a form of speech recognition as part of the verification decision.

2.6.3.1. Segmental analysis

In the segmental approach, only the relevant phonetic segments (such as particular vowels or consonants) need to be present in both the enrolment and the test utterances, and other parts of the text may be varied; thus it has been labelled *semi-text-dependent* (Sutherland and Jack 1988). Parameter measurements tend to be averaged over the segment of interest (e.g. Glenn and Kleiner 1968) or taken at only a single point (e.g. Luck 1969), so that temporal alignment is not required.

The accuracy of this method depends on correct recognition and isolation of the segments of interest. In many studies (e.g. Frederico et al. 1987) this has been done by hand, but automatic methods have been used in some cases (e.g. Fakotakis and Kokkinakis 1985, who located vowels using energy maxima), and these would be essential in a working system (see Chapter Eight, 8.4.1).

Segmental analysis has particular advantages for verification, where the text can be chosen to include segments of interest, but varied sufficiently to prevent the use of taped recordings of authorised speakers; texts can also be extremely short, reducing access and processing time. Most studies using segmental analysis have been identification studies, however (e.g. Glenn and Kleiner 1968, using filter-bank spectra of nasal stops; Goldstein 1976, using formant features from diphthongs, vowels and retroflex sounds; Sambur 1975, using formant frequencies in vowels, nasal stops and voiceless fricatives; Paul et al. 1975, using a variety of spectral features; Höfker 1977, using a variety of segments in German; and Pal and Majumder 1977, using formant frequencies in Telugu vowels). These indicate the interest in semi-text-dependent methods for applications where the text cannot be controlled, but there is a reasonable assurance that certain segments will occur often enough to be analyzed.

2.6.3.2. Vector Quantisation

The second variation, which combines elements of both the text-dependent contour-based approach and the segmental approach, is *Vector Quantisation* (Gray 1984). This is a form of clustering analysis used for data reduction, in which analysis vectors from individual speech frames throughout an utterance are classified into a relatively small number of groups (64 or 128, for example). Each group is then represented by its mean vector or *centroid*, and this group of centroid vectors is known as a Vector Quantisation *codebook*. This codebook is a highly economical representation of the entire range of analysis vectors (e.g. spectral sections or LPC parameters) occurring in the speech from which it was

prepared, and its composition is largely independent of the textual content. Its use in speaker verification has been pioneered by Soong et al. (1985) and Buck et al. (1985). In speaker verification, a codebook is formed during enrolment to characterise the reference utterances for a given speaker. During a bid, the test utterance provided is "encoded" (that is, all the parameter vectors it contains are classified) using the codebook of the claimed speaker: each analysis vector (one for each speech "frame") is compared with every codebook centroid vector, and assigned to the cluster with which it shows the greatest similarity (or the smallest distance). The distances given by each input vector are accumulated over the entire test utterance and averaged to give an overall indication of the similarity of the test utterance to the reference codebook. This distance or *distortion* measure is then passed to the classifier in the same way as any other distance measure.

This method is similar in concept to the contour and segmental analysis methods, since individual time points of the test utterance are matched against corresponding reference materials, but its great advantage is that no temporal alignment of test and reference materials is needed, since individual analysis vectors are compared with every possible reference vector (represented by the limited set of centroids) until the best match is found; thus text-independent operation is possible, if the reference and test materials are sufficiently diverse in their segmental content.

2.6.3.3. Hidden Markov Modelling

Hidden Markov Modelling or HMM is a powerful pattern recognition technique much used in speech recognition, but only recently applied to speaker verification. In its typical application, individual segment types located in the training data are used to train a set of models which predict the likelihood of a given *observation vector* (such as a cepstrum or LPC vector) originating from one of those segment types. In speech recognition, individual analysis frames can thus be assigned to the segment type with the greatest likelihood score. In speaker verification, the likelihood scores obtained by classifying all the vectors in a bid utterance using the reference speaker's models can be combined into an overall likelihood, and compared against a threshold.

The technique has been used by Poritz (1982), Tishby (1988), Zheng and Yuan (1988) for identification, Rosenberg and co-workers (1990) and Savic and Gupta (1990). Rosenberg and co-workers (1990) used sets of phoneme-based or acoustic segment-based models on a twenty-speaker isolated digit database, while Savic and Gupta (1990) used a smaller set (just five) of *broad-class* models representing nasal stops, fricatives, plosives and two classes of voiced segment.

Like the VQ approach, HMM is capable of text-independent operation if the training data encompasses enough tokens of all segment types.

2.7. Parameterisation

At the preprocessing stage, the sound wave travelling from the speaker's lips and nostrils is converted into an electrical waveform (a continuously varying voltage) by a microphone. This analogue signal is then converted to a parametric representation – that is, it is expressed in terms of variations over time in separate acoustic parameters such as signal amplitude and fundamental frequency (*time-domain* parameters), or distribution of spectral energy (*frequency-domain* parameters). This "parameterisation" may be done by analogue methods (e.g. the use of a bank of electrical filter circuits, responding to different frequency bands, to produce one form of spectral representation), or it may be performed digitally by sampling the electrical waveform and submitting this digitised speech signal to a variety of computer-based signal processing techniques (including fundamental frequency tracking, Fourier analysis or Linear Predictive Coding). Digital techniques are generally based on analysis of a certain length of speech data (an analysis *frame*), so the resulting parametric representation is actually discrete, with values for each parameter at successive instants throughout the utterance; analogue methods produce parameters which vary continuously in time in the same way as (though usually at a lower rate than) the original speech, and these parameters must themselves be sampled at suitable intervals to produce the discrete representation required for the formation of a feature vector.

The most commonly used methods of parameterisation yield information on just three aspects of speech: the energy present in the speech waveform, the

fundamental frequency of vibration of the vocal folds and the shape of the speech spectrum. These parameters are sufficient to allow a reconstruction (synthesis) of almost any section of a speech waveform, and are widely used in other forms of speech processing (speech recognition, speech storage and speech transmission, for example). Their use in verification studies in many cases stems from a desire to avoid additional processing. Almost all verification studies use one or more of these three types of parameter, in fact.

2.7.1. Gain or intensity

Gain and intensity are used interchangeably in the literature to refer to measures of the sound energy present in the speech waveform. The energy in a wave is related both to the amplitude of vibration of the wave and to the frequency with which the vibration is taking place. The amount of energy in the speech waveform over a given period of time depends mainly on i) the sub-glottal pressure (determining the amplitude of vibration of the vocal folds, and to a lesser extent the fundamental frequency); and ii) the degree of opening (or the degree of constriction) of the vocal tract. Variation in either of these factors will cause corresponding variation in the intensity contour measured during speech. For example, vowels generally have a more open vocal tract than other segments, and so are usually more intense; open vowels such as [aa] will, other things being equal, have greater intensity than close vowels. The intensity contour is therefore highly dependent on the segmental content of the utterance. Factors affecting sub-glottal pressure include rhythmic word accent and nuclear stress, the addition of particular emphasis to a word, and the

speaker's general disposition. Markel, Oshika and Gray (1977) suggest that the amount of intensity variation over time is correlated with different perceived voice characteristics: speakers whose voices are judged "emphatic", "dynamic" or "lively" will show greater intensity variation than those with "flat", "boring" or "monotonous" voices. There has been some interest, then, in the measurement of intensity for speaker recognition.

2.7.1.1. Measuring intensity

Speech intensity, according to Fant (1970, p.229) is almost never measured directly, but is derived from measures of the amplitude of the sound wave over a short time interval (the *integration* or *averaging* time: Fant 1970). There are several different methods in use, but all produce very similar parameters. Atal (1976) recommends integration (or summation) of the square of the amplitude values over a time interval of between 10 and 30 milliseconds. Squaring the values allows the negative half of the speech waveform to contribute to the final measurement, and also reflects the fact that intensity is proportional to the square of the amplitude for a given frequency (Fry 1979). This was the method used by Markel, Oshika and Gray (1977). Wasson and Donaldson (1975) calculated the mean of the absolute values of the amplitude of the speech wave over an interval of 10 msecs. Other possibilities are the rms (root-mean-square) value over the integration time, and (the crudest of all) the simple amplitude envelope obtained from the points of maximum displacement (Doddington 1976). This method fails to take frequency variations into account.

It is also possible to obtain intensity information from spectral representations of the speech waveform, over a similar time interval. Lummis and Rosenberg (1972) and Lummis (1973), for example, derived their gain measurement by summing the amplitude values in the low-frequency "bins" (from 0 to 600 Hz) of the FFT spectrum, a process equivalent to integration of the spectral curve. Other spectral analysis algorithms produce an estimate of the gain automatically: the zeroth autocorrelation coefficient, for example, derived during Linear Predictive Coding, is simply the sum of the squared waveform values during the analysis frame.

Use of intensity

Intensity or gain has been used for speaker verification but always in conjunction with another parameter such as fundamental frequency. Both gain *contours* (over short, fixed text utterances) and gain *statistics* (over longer utterances, usually free-text) have been considered. In both cases some form of normalisation is required, so that overall differences in the intensity of the speech signal resulting from different mouth-to-microphone distances or attenuation settings in recording equipment do not create false differences between speakers.

Gain *contours* have been used in conjunction with fundamental frequency and formant contours by Doddington (1971), Lummis and Rosenberg (1972) and Lummis (1973); in conjunction with fundamental frequency and LPC coefficient contours by Rosenberg and Sambur (1975); and with fundamental frequency only by McGonegal, Rosenberg and Rabiner (1979). In Lummis' study (1973),

the gain contour gave the lowest error rate of all individual parameters (1.7% average EER), but the fact that Doddington omits it from his later work (1976) suggests that it was actually of limited value.

Statistical measures of gain variation have been examined by Wasson and Donaldson (1975); Markel, Oshika and Gray (1977); Markel and Davis (1979); and Johnson, Hollien and Hicks (1984). Wasson and Donaldson (1975) used the average speech amplitude in conjunction with two measures of zero-crossing rate in short sentences. Markel, Oshika and Gray (1977) used the dispersion (defined as the standard deviation divided by the mean) as a measure of long-term variation in conversational speech, along with statistics of fundamental frequency and LPC reflection coefficients. In their evaluation, the gain parameter was the least effective. Nevertheless, Markel and Davis (1979) retained gain measurement (mean, standard deviation and dispersion) in their study of speech from interviews. Johnson, Hollien and Hicks (1984) used statistics of the distribution of energy in a number of amplitude levels (relative to the peak amplitude during the utterance), along with information on the temporal distribution of voicing. More details of this last study are given in section 2.7.4.

2.7.2. Parameters describing phonation

2.7.2.1. Mechanism and importance of phonation

The source of acoustic energy for speech is, for the most part, a periodic acoustic waveform produced by the passive vibration of the vocal folds (*voicing* or *phonation*). The frequency with which the vocal folds vibrate is determined

at any instant by the tension under which they are held, their mass and intrinsic mobility, and the pressure produced in the trachea by the expulsion of air from the lungs. It is this vibration which imposes a periodicity or regularity on the acoustic waveform. The frequency of repetition of the acoustic waveform is known as the fundamental frequency, and corresponds closely to the frequency of vibration of the vocal folds. The duration of a single cycle of vocal fold vibration is referred to as a pitch period: the greater the frequency of vibration, the shorter the period.

The regularity of vibration of the vocal folds is not perfect, however: the duration of successive pitch periods, and consequently of the measured fundamental frequency over a given interval, can vary owing to factors both within and outwith the speaker's control.

Fundamental frequency is controlled for linguistic purposes: variation in fundamental frequency forms the chief basis of patterns of intonation, and frequently of lexical stress or accent. It also reflects social and personal attributes of the speaker. In particular, the habitual range of fundamental frequencies used by a speaker tends to reflect factors such as body size and build: speakers with a larger larynx, and (presumably) more massive vocal folds, tend to use lower fundamental frequency ranges than those with a smaller larynx. The most obvious manifestation of this is the difference in fundamental frequencies shown by the sexes. These physically-related differences can be modified, however, by social factors – certain accent groups, for example, may favour a higher fundamental frequency in male speakers than others – and by physio-

logical change: the production of mucus on the surface of the vocal folds or a build-up of fluid in their tissues, for instance, will lead to an increase in the mass of the vocal folds and a lowering of the frequency with which they will vibrate.

Whatever their cause, though, differences in fundamental frequency between different speakers and groups of speakers have been seen as having great potential for automatic speaker recognition, and have indeed been shown to play a major role in aiding naive speaker recognition (e.g. Brown 1980). Part of their attraction is that a wide range of analysis techniques now exists for the measurement of fundamental frequency. Such methods are often termed "Pitch Determination Algorithms" or PDAs, *pitch* being used interchangeably with fundamental frequency throughout the literature, despite the fact that, as a technical term, it refers to its perceptual correlate. An extensive review of PDAs is found in Hess (1983). Several spectral analysis techniques, such as cepstral analysis and Linear Prediction, can also provide estimates of fundamental frequency, and it is usually these methods which have been used in verification studies, since spectral parameters can be derived at the same time: Furui (1981b), for example, derived fundamental frequency from the peak position of the correlation function of the prediction residual produced during the calculation of the PARCOR coefficients (the precursors of the Log Area Ratios actually used in his study).

Other aspects of phonation, such as phonation type ("voice quality", 2.7.2.3) and temporal aspects of voicing (2.7.2.4), have also been considered,

though not to the same extent.

2.7.2.2. Fundamental frequency

Fundamental frequency parameters have been described using characterisations of contours, statistical distributions over time and isolated values taken in single segments. Most studies use fundamental frequency in combination with another parameter (gain or spectral data), the fundamental frequency information being a by-product of the spectral analysis itself.

Fundamental frequency contours have been used for text-dependent verification by Doddington (1971), Lummis (1973), Rosenberg and Sambur (1975), McGonegal, Rosenberg and Rabiner (1979), Sambur (1975) and Furui (1981b). Lummis (1973) achieved an average equal error rate over all speakers of 4.1%. McGonegal, Rosenberg and Rabiner (1979) looked at the robustness of pitch (and gain) contours alone in the face of alterations to the transmission system used for telephone channels between training and testing a system. No significant effects on verification error resulted from differences in the processing system used, even when this was different between training and test data. Error rates varied greatly from speaker to speaker, however, and the sexes performed differently, with the median average verification error (that is, the median value of all speakers' average error $FR + FA/2$) being 12% for males and 8% for females. Sambur (1975) examined fundamental frequency features in the single sentence "*Cash this bond, please*", using the zero-crossing rate in the low-pass filtered speech waveform. The features comprised the slope of the contour from the onset of voicing to the peak fundamental frequency in "*cash*"; the

slope from this peak to the end of the vowel; the slope during the vowel in "*bond*"; the mean fundamental frequency in each voiced section of the sentence; and the overall mean fundamental frequency. This last feature had the highest place of all phonation parameters in the ranking by the "probability of error" measure devised by Sambur, but came only twelfth out of all features examined. Slope features were not very effective. Sambur noted that speakers' fundamental frequency varied markedly from session to session. No details are given of the performance of the fundamental frequency parameters in Doddington (1971) or Furui (1981b).

Fundamental frequency measurements from individual segments have been used by Luck (1969), Frederico and co-workers (1987), Paoloni and co-workers (1986), Wolf (1972), and Paul and co-workers (1975). Only Wolf gives details of the performance of fundamental frequency independently of the other parameters used, however. Fundamental frequency was estimated from the zero-crossing rate of the low-pass filtered waveform at fixed points in six sentences (usually at the middle of stressed vowels). These measurements occupied the first nine places in an F-ratio ranking of all the features examined (including nasal formant frequencies). An attempt to measure the increment in fundamental frequency accompanying lexical stress (by subtracting the F0 of preceding unstressed vowels) was abandoned owing to a rather high intraspeaker variance. The F0 features showed high degrees of intercorrelation, and it appears that only one was selected for the verification and identification trials.

Fundamental frequency statistics have been widely used, and are common in text-independent studies (mostly for speaker identification), but again their performance is often not indicated. Das and Mohn (1971), for example, used means and variances calculated over two short intervals in the phrase "*Check available terminals*".The two means (but not the variances) were ranked among the top 200 features by F-ratio, but no information is given on their individual effectiveness. Clarke and Becker (1969), who also used the mean and variance calculated over short sentences, give only identification results: mean F0 gave an accuracy of 42%, with variability measures reaching between 32% and 35%; these compare with the 63% achieved using the long-term average spectrum. Markel and Davis (1979) achieved an average verification error of 4.25% on free-text interviews using fundamental frequency means and standard deviations with ten reflection coefficients. *Identification* studies using F0 statistics include Mead (1974), Hunt, Yates and Bridle (1977), Hunt (1983), Hollien, Johnson and Doherty (1978) and Green (1972).

Some attempts have been made to use statistics relating to features such as rates of change in fundamental frequency contours, and contour slope; this sort of approach preserves some of the temporal information normally lost in statistical studies, but allows text-independence. Hunt, Yates and Bridle (1977) examined various statistical representations of fundamental frequency behaviour in short speech samples (extracts from radio broadcasts, twenty or thirty seconds long), for text-independent identification using a real-time cepstrum processor. The parameters investigated comprised the mean, mean devi-

ation and low-order moments of the distribution about the mean of the fundamental frequency and its first and second-order difference curves; parameters related to the correlation between fundamental frequency and its rate of change; parameters relating to the proportion of time the fundamental was rising or falling; and statistics of sections of F0 curves divided up according to their slope. The simplest statistics were found to be the most useful, however, with the overall mean fundamental frequency the best of all. The best six parameters – the mean, mean deviation, second and third moments about the mean (variance and skewness), the proportion of time F0 was falling, and a "signed" second moment (apparently the sum of squared deviations from the mean, with the sign of the deviation restored) – were then included in a set of 46 parameters (the other forty relating to the distributions of the first 8 cepstral coefficients) in a validation experiment using a separate data set. The six as a group performed better than the statistics of any one cepstral coefficient, but this is attributed to the fact that the cepstral means could not be used because they are dependent on channel characteristics. Fundamental frequency parameters appeared to be independent of cepstral parameters.

Fundamental frequency on its own appears to be of limited use, then. Several studies have found it to be inferior to spectral parameters, and less stable over time (Markel and Davis 1979), but because it is independent of spectral information (Hunt 1977), it can be useful to include it in a combined feature set (e.g. Cheung and Eisenstein 1978).

2.7.2.3. Phonatory quality and perturbation analysis

The periodicity of the source function – the glottal waveform – is not its only characteristic. There is also some variation in the fine detail of the *shape* of the waveform, and especially in its regularity (in both frequency and amplitude of vibration). This variation is caused by a variety of factors, such as asymmetry in the structure of the vocal folds and variations in supraglottal air pressure. Some of the consequences of this variation will be visible in the time-domain, particularly in the regularity of the durations of individual periods or glottal cycles. Others will be seen in the frequency domain, since the temporal quality of the glottal wave is reflected in its frequency spectrum, which forms the input to the modifying vocal tract filter.

The spectrum of the excitation function or glottal source has been studied using the technique of inverse filtering - that is, passing a speech sample through a filter whose frequency response is the inverse of the estimated vocal tract transfer function. Examples of such studies are Carr and Trill (1964) and Mártony (1965). Although the possibility of using source spectra for speaker recognition is mentioned by Carr and Trill (p. 2036), no published study has used this particular technique. Wolf (1972), however, used a crude estimate of the slope of the source spectrum in his verification study. The feature used was the difference (in dB) between the amplitudes of the first and third formants of the vowel [uu], after normalisation by their separation on a logarithmic frequency scale. This feature came fourth in the ranking by F-ratio (after the mean fundamental frequency and the formant frequencies of nasals), and

showed very little correlation with other spectral features (as indicated by Wolf's ΔP measure of inter-parameter dependence).

It should be noted here that Nolan (1983) demonstrates that the long-term average speech spectrum (see section 2.7.3), widely regarded as conveying information on the effects of supralaryngeal structure and behaviour, shows remarkably little susceptibility to gross changes in supralaryngeal setting, but considerable susceptibility to changes in phonatory quality instead. It could be argued, therefore – and is argued by Nolan himself (195) –, that the use of the long-term average spectrum is already tapping qualities of the laryngeal source. The long-term spectrum certainly must contain *some* source information, since it is virtually impossible to separate the source and filtering characteristics of the vocal tract completely.

2.7.2.4. Temporal distribution of voicing

A very few studies have considered the temporal distribution of phonation in speech.

Wolf (1972) included a feature indicating the frequency with which speakers prevoiced a phonemically voiced stop when it followed a voiceless fricative after a word boundary (/b/ in the phrase "*Cash this bond, please*"). Prevoicing of phonemically voiced stops is not required in English in any context, and its frequency varies from speaker to speaker. Wolf classed a stop as prevoiced if voicing preceded the stop release by 20 milliseconds or more. Each speaker produced ten tokens of the relevant sequence. Six of 21 speakers showed prevoicing of the stop more than half the time; 4 showed it occasionally; while the

remaining 11 speakers showed none. Wolf found the feature appealing, despite its crudity and limited value (it came rather low in the ranking by F-ratio), because

> it concerns a rapid event which is not likely to be consciously modifiable and .. is an event of such specificity that it is probably independent of most other parameters"

(1972: 2051).

Sambur (1975), using the same speech database (though extended to cover over three and a half years), examined another aspect of voicing onset in stops: the duration of the frication and aspiration noise (that is, the voicing lag) in an initial voiceless stop [k]. This feature was measured manually from waveforms of speakers' utterances of the same sentence "*Cash this bond, please*".It was ranked fourth by Sambur's "probability of error" criterion.

Johnson, Hollien and Hicks (1984) looked at voicing behaviour over a much longer time, along with the temporal distribution of energy in the speech waveform. The parameter used was simply the total time during which voicing was present (presumably expressed as a fraction of the total speaking time), together with the total articulation time itself. This two dimensional vector was not as effective as the time-energy distribution vector examined at the same time (see section 2.7.4).

Nolan (1983: 129) suggests alternative measures of the temporal distribution of voicing, such as:

> mean durations of the alternately voiced and voiceless stretches (preferably excluding pauses) of which the speech wave consists; the ratio between these means; and statistical distributions of each kind of stretch

equating them with perceptual contrasts such as Crystal's *spiky/glissando* (Crystal 1969) and the everyday terms *clipped/drawled*:

> "clipped" speech would consist of relatively short (principally vocalic) stretches and relatively long intervening voiceless (principally consonantal) stretches, whilst "drawled" speech would be perceived when the opposite relationship existed.

(1983: 129). He points out, though that such parameters would be especially affected by speakers' communicative intent. In addition, they would be particularly prone to imitation by impostors, like fundamental frequency, and have never been used in verification studies.

2.7.3. Vocal tract parameters: the speech spectrum

The majority of studies in verification use spectral parameters, which describe the distribution of energy in the speech waveform over the range of frequencies. The output spectrum of speech derives its shape from the filtering action of the vocal tract, acting on the spectrum of the glottal sound wave. Since the filtering characteristics of the vocal tract are determined by the dimensions of the vocal tract cavities and the position of the articulators, the speech spectrum contains a large amount of information suitable for verification. It also contains much information about the segmental content of speech, but in most cases no attempt is made to separate the two types of information.

A wide range of parameterisation methods is available. Methods used in verification include *filter-bank* analysis, Fourier analysis in the shape of the *Discrete Fourier Transform*, *Cepstrum* analysis and – possibly the most power-

ful and most widely used technique – *Linear Predictive Coding*.

The studies fall into two main categories: those which use the overall distribution of energy in the spectrum directly, and those which extract *resonance* frequencies from the spectral representation. The overall spectral energy can be represented by the frequency spectrum itself, by the coefficients of the cepstrum, or by any one of the coefficient sets produced by Linear Prediction; the feature vectors used in this approach therefore tend to be rather large, and the number of samples required for stability is quite high. The extraction of resonance features from the spectrum requires an extra processing step which is prone to error, but produces vectors of lower dimensionality. Features are typically *formant* frequencies, representing the natural resonant frequencies of the vocal tract, and can be extracted by locating the peaks in the power spectrum or by polynomial root extraction (Markel 1972) on cepstral or Linear Prediction coefficients.

2.7.3.1. Filter Bank Analysis

A filter bank is a set of band-pass filters spanning a particular frequency range. Each filter is designed to respond to a different band of frequencies, and only energy present in that frequency band will be represented in the waveform of its output. The output of each filter is measured at discrete intervals of time (sampled), usually by rectification and integration over a short interval (e.g. 10 millisecs). Over any one interval, then, the distribution of energy across the frequency bands of the filter bank constitutes a short-time speech spectrum.

Filter bank analysis (including spectrography) has been widely used in speaker verification. Studies which have used the whole spectrum include Li et al. (1966), Doddington (1976, 1985), and Ney and Gierloff (1982) (contour-based); Das and Mohn (1971) (segmental); and Bunge (1977a) and Mohankrishnan et al. (1982) (statistical).

In Doddington's system (1976, 1985), which controls access to the Texas Instruments Computer Centre in Dallas, only frames from the centre of the vowel in each word are used, with time alignment being achieved by choosing those frames which give the minimum distance from the reference template. Each verification utterance consists of four words, built up randomly from a set of sixteen, and the distances for each word are combined to give a final distance for the verification decision. This system has been quite successful: trials on 200 users over several years have given error rates below 1%, using a sequential decision strategy in which up to four phrases can be used in each attempt, and up to two attempts are allowed.

Ney and Gierloff (1982) used telephone recordings of a single fixed sentence, and normalised the output spectra by their long-term average to take account of varying transmission line characteristics; groups of adjacent spectra were averaged to reduce each data matrix or template to exactly 64 spectral vectors. An average error rate of between 1% and 1.5% was obtained.

Bunge (1977a) used the long-term average spectrum derived from a 43-channel filter-bank for text-dependent and text-independent verification, achieving an equal error rate of 1% in both cases on a relatively large database

(41 male and 9 female speakers, each providing two sets of 50 utterances). Mohankrishnan et al. (1982) considered the long-term average of a 16-channel filter-bank along with Linear Prediction parameters; the best results (0.91% False Acceptance and 1.48% False Rejection) were achieved using both the filter-bank spectrum and the (Linear Prediction) Inverse Filter spectrum.

A few studies have extracted formant-based features from filter-bank spectra, including Wolf (1972) (segmental), Feix and De George (1985) (contour-based) and Das and Mohn (1971) (statistical). Extraction of resonance features has been attempted in various ways. Feix and De George (1985) generated a set of 32 *binary* valued features from a 16-channel digital filter-bank, perceptually scaled to emphasize formant information. These features included the presence of maxima in selected spectral bands, spectral slope features at particular frequencies and unspecified "vowel class" features. In trials with 18 speakers over five months, rates of 0.6% False Rejection and 0.3% False Acceptance were achieved. Das and Mohn (1971) extracted both spectral energies and formant features from a filterbank analysis of selected segments in the phrase "*Check available terminals*".A set of 405 features was examined, comprising the frequencies and amplitudes of the first three formants, the spacing between them, formant slopes, spectral averages in different frequency bands over different lengths of time, durational features, and some statistics of the fundamental frequency. The best 200 features gave a mean error rate of 1%, but the system failed to decide in 10% of cases (because of segmentation errors, for example).

Wolf (1972) used a 30-channel analogue filter bank analysis as the basis for a variety of segment-based spectral and temporal features, in an attempt to find an efficient feature set for verification and identification. Only single segments, taken from particular points in six fixed sentences, were analysed. These segments comprised the nasal stops [m] and [n], the vowels [aa], [a], [@] and [ii], and the fricative [sh]. With the nasals it was not possible to locate the poles and zeroes in the filter-bank spectrum, so the amplitude values of individual filter outputs (normalised for the amplitude of the following vowel) were used instead: the filter outputs chosen (by F-ratio) corresponded to the nasal resonance peaks at 250 Hz, 2000 Hz and 3000 Hz in [m], and those at 1000, 2000 and 3000 Hz in [n]. Formant features for the vowels were estimated in three different ways: analysis-by-synthesis was used to estimate F1 and F2 for [a] and [aa]; direct estimation from the filter-bank spectrum gave F1 and F2 for [@]; while for [ii] and [aa], statistical descriptions of the spread and skewness of spectral energy over particular formant regions (F2 to F4 for [ii], F1 to F2 for [aa]) were used, to avoid the problem of separating formant peaks. Other features used included an estimate of the voice spectrum slope taken from the vowel [uu]; and a gross categorisation of the shape of the high frequency spectrum during [sh]. Also considered were fundamental frequency at selected points in two of the sentences; the duration of a single word ("*bought*"); and the presence or absence (a binary feature) of prevoicing in /b/ after word-final /s/. Parameters were selected on the basis of their F-ratios, with inter-parameter correlations also being taken into account. An EER of 2% was achieved with 17 parameters.

2.7.3.2. The Discrete Fourier Transform

Fourier analysis, in the shape of the Discrete Fourier Transform or DFT, provides an alternative method of deriving the speech spectrum for a given analysis "window", but is not much used in verification studies. The preferred use of the DFT is in the calculation of the cepstrum (see below).

The amplitude of the spectral components in a finite-length sequence of speech samples is evaluated at discrete frequency intervals over the range 0 Hz to $F/2$ Hz, (where F is the sampling frequency). The result is a set of *Fourier coefficients*, comprising the *amplitude* and *phase* of each spectral component. Phase information is predictable from the amplitude spectrum, and is frequently considered to contribute little to the *perceived* quality of the sound (see the discussion by Cox and Robinson (1980)). For these reasons, phase information is usually discarded and only the amplitude spectrum considered (Fant 1970). The envelope of the spectral components for a given analysis frame provides a similar representation of the speech spectrum to that given by filter-bank analysis, but can give greater detail; however, some form of smoothing (e.g. by cepstral analysis) is usually needed if formant frequencies are to be extracted. The DFT of a speech signal is usually derived indirectly, by means of a set of algorithms known as the *Fast Fourier Transform* or FFT (Atal 1985), a fast and computationally efficient method of calculating the Fourier coefficients.

Only one verification study has used Fourier analysis to provide spectral parameters. Luck (1969) used the Fourier coefficients themselves to character-

ise two vowel segments (/ai/ and /ou/) in the fixed sentence "*My code is ...*", in combination with 16 cepstral coefficients, the fundamental frequency and a duration parameter. However, two identification studies have used Fourier spectra. Bogner (1981) found that a low-pass filtered version of the spectrum – the Filtered Logarithmic Spectrum, produced by logging, differentiating and smoothing the power spectrum – was resistant to changes in transmission line effects, while the Linear Prediction reflection coefficients were not. Paul and co-workers (1975) used FFT spectra along with LPC-derived spectra in semi-automatic speaker identification based on isolation of vowels and nasal stops, but no comparison is given.

2.7.3.3. Cepstral Analysis

Cepstral analysis provides a method of separating the source characteristics (that is, the effects of the periodicity of the excitation waveform from the vocal folds) from the vocal tract filter characteristics. Examination of the short-time speech spectrum shows that the spacing of the harmonics of the fundamental frequency in the glottal source spectrum produces a ripple on the output speech spectrum. The "wavelength" of this ripple is equal to the harmonic interval, and therefore (since harmonics are present in the glottal spectrum at integer multiples of the fundamental frequency) to the fundamental frequency itself. The convolution of the two spectra – the rapidly varying excitation spectrum and the smooth, slowly varying vocal tract spectrum – can be overcome by taking a Fourier transform of the logarithm of the power spectrum. This yields a function known as the *cepstrum*, a pseudo-time function in which

the slowly-varying vocal tract function is represented by cepstral components near the x-axis (or "quefrency") origin, while the periodicity of the excitation is represented by a large peak along the x-axis at a distance equal to the period (in milliseconds) of the glottal cycle (that is, the inverse of the fundamental frequency) (Noll 1967). An inverse Fourier transform carried out on the low-quefrency components (below the pitch-period peak) produces a smoothed version of the speech spectrum, containing predominantly the contribution of the transfer function of the vocal tract, with the disturbing ripple of the excitation function removed (Schafer and Rabiner 1970). Thus cepstral analysis can provide either spectral parameters (in the form of the coefficients of the smoothed spectrum or the low-quefrency cepstral coefficients themselves), or an estimate of pitch period duration (used by Doddington 1971, Lummis 1973, Hunt 1983).

The cepstral coefficients themselves were used by Luck (1969), in a segment-based study which also included the use of Fourier coefficients. Hunt and co-workers (1977), in an identification study, used a real-time cepstrum processor to derive statistics of the fundamental frequency and the first 8 cepstral coefficients from unconstrained FM radio broadcasts; the independence of the cepstrum and the pitch parameters was demonstrated by the fact that the error rate for the combined set was the product of the two individual error rates.

Cepstral processing has also been used to derive spectral formant frequencies. Lummis (1973) used the cepstrally smoothed spectrum to find formant frequencies in a single sentence, but the resulting contours were not as effective

as either the fundamental frequency or the gain; while Frederico and co-workers (1987) derived formant frequencies from vowels in continuous, unconstrained speech by polynomial root extraction from the cepstrum.

2.7.3.4. Linear Predictive Coding (LPC)

Linear Predictive Coding is a time-domain method in which the amplitude of a sample point is estimated ("predicted") by a linear combination of its past values, weighted with appropriate predictor coefficients, and its present value (see Chapter Four, 4.2). The values of the predictor coefficients are chosen so as to minimise the mean squared error value between the predicted waveform and the original signal. If this is done well, the spectrum of the error signal – or the *residual* as it is known – approaches that of white noise (that is, it has equal amplitude at all frequencies). The resulting predictor coefficients form a model of the speech waveform which combines the spectral effects of the vocal tract filter with the contributions of the glottal flow within a pitch period and the effects of lip radiation (Atal 1985). The effects of the periodic excitation of the vocal tract by the vocal folds are removed. This model can be viewed as a filter, whose transfer function is the inverse of the vocal tract transfer function (along with the effects of the glottal source spectrum and lip radiation). Thus Linear Prediction provides an efficient way of representing the short-time speech spectrum, with the added advantage that the rippling effect of the periodic excitation is removed.

The Linear Prediction model is an all-pole model, however, and assumes that the vocal tract has no side-branches contributing zeroes to its transfer

function. It therefore fails to model the speech spectrum accurately during segments such as nasal stops, when the mouth forms a side-branch to the nasal-pharyngeal "tube", and nasalised vowels, when the nasal cavities constitute the side branch (see Chapter Three, 3.4.1).

Parameterisation by Linear Prediction yields either a set of predictor coefficients, then, or an estimate of the short-time spectrum. The predictor coefficients themselves are fewer in number than the spectral coefficients, and therefore give vectors of lower dimensionality; they have been used as contours by Atal (1974) and Rosenberg and Sambur (1975); in Vector Quantisation by Soong and co-workers (1985); and in statistical analysis for text-independent operation by Mohankrishnan and co-workers (1982), Sambur (1976), Everett (1985), Li and Wrench (1983) and Chen and Lin (1987). Li and Wrench used a statistical characterisation of speakers' VQ distortion measures for speaker identification.

The Linear Prediction spectrum has been used by Naik and Doddington (1986, 1987), Nolan (1983), Mohankrishan and co-workers (1982), Pfeifer (1974), Sambur (1976), Paul and co-workers (1975) and Fakotakis and Kokkinakis (1985). Naik and Doddington converted the spectrum into a set of amplitude values simulating the output of a 14-channel mel-based filter-bank (thus reducing its dimensionality), and transformed the values using Principal Components analysis. Mohankrishnan and co-workers found that the Linear Prediction inverse filter spectrum was the best representation among those examined (predictor coefficients, reflection coefficients, Log Area Ratios, cep-

strum coefficients and a filter-bank spectrum), giving an Equal Error Rate of 2.47%. Pfeifer (1974), in an identification study, also found that the Linear Prediction spectrum was superior to the inverse filter coefficients themselves and the reflection coefficients. It is possible that this is because its high dimensionality makes the resulting error rate rather optimistic, however. Paul and co-workers (1975) and Fakotakis and Kokkinakis (1985) used the Linear Prediction spectrum to obtain formant frequencies, both in segments extracted from sentences.

It is also possible to use the residual signal for pitch extraction since it is almost never completely flat, but shows some correlation with the periodicity of the original signal. Examples of its use are McGonegal et al. (1979) and Furui (1981b).

Other parameter sets related to the predictor coefficients are frequently used in speech analysis and particularly in speaker verification. The autocorrelation method of Linear Prediction (Atal 1985) yields a set of *autocorrelation* coefficients which can be used in their own right, without the need to derive predictor coefficients from them; autocorrelation coefficients have been used by Atal (1974) in a comparative study of Linear Prediction parameters, and by Rosenberg and Shipley (1983), who used just 8 coefficient contours for verification as part of an isolated word recognition system. Similarly, the *Partial Correlation* coefficients (PARCOR coefficients) of the speech signal also show a direct relationship with the predictor coefficients, and can be used as parameters in their stead (e.g. Guo et al. 1987, Sambur 1976, Chen and Lin

1987). The PARCOR coefficients are also directly related to the *reflection coefficients* of an acoustic tube model of the vocal tract, in which the (all-pole) transfer function is modelled by a series of lossless acoustic tubes of equal length but varying cylindrical cross-section; reflection coefficients have been used by Mohankrishnan and co-workers (1982), Pfeifer (1974), Markel and Davis (1979), Schwartz and co-workers (1982), Bogner (1981) and Savic and Gupta (1990). The relationships between the differing cross-sectional areas of the tubes in this model provide another representation – the *Log Area Ratios* or LARs – used by Mohankrishnan and co-workers (1982), Sambur (1976), Furui (1981b), Schwartz and co-workers (1982), Wolf and co-workers (1983) Krasner and co-workers (1984) and Savic and Gupta (1990). In Sambur's study, the Log Area Ratios performed as well as the PARCOR coefficients, and better than the predictor coefficients themselves.

A final alternative representation is given by the derivation of the *cepstrum* of the transfer function: this can be done directly from the predictor coefficients, without the need to return to the frequency domain, and the resulting *complex cepstral coefficients* are frequently used in speaker recognition studies, such as Atal (1974), Furui (1981a), Soong and Rosenberg (1986), Rosenberg and Soong (1986), Higgins and Wohlford (1986), Mohankrishnan and co-workers (1982), Schwartz and co-workers (1982), Krasner and co-workers (1983), Gish and co-workers (1985, 1986), Rosenberg and co-workers (1990) and Savic and Gupta (1990). Atal (1974) and Furui (1981a) found them to be superior to other Linear Prediction representations, though Mohankrishan and co-

workers obtained better results with the Linear Prediction spectrum. One advantage of using the LP cepstrum is the speed with which it can be computed: Furui points out that the FFT-derived cepstrum gives similar results to the Linear Prediction cepstrum, but takes twice as long to derive. Another attraction is that the effects of transmission channel variation can be removed by subtracting the mean cepstral vector from all utterances (Furui 1981a, Gish et al. 1985, 1986). Schwartz and co-workers, however, found that they performed badly on noisy speech of telephone bandwidth.

2.8. Temporal features

Temporal features have received less attention than any other type of feature in the speaker recognition literature, partly because there are so many poorly understood linguistic constraints and influences on the time-course of speech. While they have the advantage over spectral parameters that they are resistant to most forms of distortion in transmission, they suffer from two major disadvantages: first, they are a likely target for mimicry by impostors; and second, their measurement is rather difficult compared with spectral parameters.

Studies which have included temporal features have tended to treat them rather crudely, and almost always as part of a wide range of non-temporal features.

The crudest temporal features are measures of the duration of whole utterances or of items (such as words or segments) excerpted from larger utterances. Clarke and Becker (1969), for example, included the overall duration of

speakers' tokens of short sentences (after normalisation for differences in sentence length) in their study, comparing it with spectral features and fundamental frequency. Using utterance duration alone gave an identification accuracy of 32%. Luck (1969) used the duration of the first word in the code phrase spoken by speakers ("*My code is* .. "), along with a measurement of fundamental frequency and cepstral measurements, but no separate assessment of its efficiency was made. Das and Mohn (1971) measured the intervals between key segmentation points in their phrase "*Check available terminals*"; both absolute durations and relative durations were used (the latter normalised for the duration of the body of the phrase itself, from /v/ to /z/). Only one interval – the final unstressed syllable of "*available*" – was retained after ranking by F-ratio.

Wolf (1972) included the duration (measured by hand and crudely quantized) of the word "*bought*" (in "*A few boys bought them*") as one of his features. This feature showed a rather low F-ratio, but also a low correlation with other features in the set examined. Like all the speech events mentioned so far, however, it seems to have been chosen more for ease of measurement (beginning and ending with a stop) than on any theoretical basis. In such a position, it would be highly susceptible to different interpretations of where the tonic stress should lie in the sentence, and would therefore have a wide range of durations across speakers, and considerable variability within speakers.

Later studies have shown slightly more sophistication in their approach to the selection of timing parameters. Sambur (1975) examined the slope of a linear (mean square) fit to the second formant in tokens of the diphthong /ai/

(manually segmented). This feature was ranked 10th (out of the 92 examined). The other temporal feature used (the voicing lag shown after the voiceless plosive /k/ in "*Cash this bond, please*") has already been mentioned in section 2.7.2.4; it was ranked fourth.

Wood (1978) included unspecified temporal features of isolated vowel, nasal and liquid segments extracted from continuous speech in a semi-automatic speaker identification system, but no details of these features are available.

A study by Johnson, Hollien and Hicks (1984) deserves special mention for its statistical characterisation of the timing of speakers' utterances. This study used two feature vectors: a Time-Energy distribution (TED) and a Voiced-Voiceless distribution (VVL). The TED vector represented the distribution of speech energy in several amplitude bands over time: the vector had 40 elements, comprising the means and standard deviations of the number and durations of energy "bursts" in each band, as well as the standard deviation of the pause periods (when no energy was present in the specified band). The VVL vector consisted of just two elements – the proportion of time occupied by voiced speech in the sample, and the total articulation time. Forty males between the ages of 25 and 45 years provided a single sample of read text lasting 2.5 minutes, under each of three conditions: with voice disguise, under stress (from randomly-administered electric shocks) and without disguise or stress ("normal"). Feature vectors were transformed using discriminant analysis. The 40-element TED vector gave better identification results than

the much smaller VVL vector under all speech conditions, though the addition of the VVL vector to TED improved its effectiveness in the disguise condition. The comparability of the two parameter sets is doubtful, though, given Foley's (1972) requirement on the ratio of sample size to vector dimensionality when training and testing are done on the same data (see 7.2.2).

2.9. Nasality in Automatic Speaker Verification

In this thesis, it has been decided to use features relating to nasality, but so far little has been said of it in this review. This section examines the case for using nasality and describes some studies in which it figures.

2.9.1. The case for nasality

The principal argument advanced in favour of the use of nasality in speaker verification (and in recognition in general) is that it provides a strong candidate for an organically-determined feature: the nasal cavities are generally recognized as being relatively fixed, unlike other vocal tract cavities, and differing widely in size from one speaker to another. This suggests that there will be consistent acoustic differences between speakers (in other words, high *between-speaker variability*). They are, however, prone to physiological changes, as will be seen in Chapter Three, and may therefore display a rather large amount of *within-speaker variability*. The nature and degree of their variability has never been assessed satisfactorily, however, particularly over a fairly long period of time in a large group of speakers.

Resistance to mimicry is another potential advantage of nasality. This stems in part from its reflection of the anatomy of speakers, but also from the fact that its spectral structure is not at all salient: it is difficult to conceive of an impostor knowing what to imitate in a speaker's nasal stops, and most unlikely that any attempt would succeed in replicating the spectrum exactly. This has not been explored in the verification literature, however.

Nasal stops have also been chosen because of their relatively high frequency in English (Glenn and Kleiner 1968); this is not a major concern in verification, however, since the speech materials can be chosen to increase the occurrence of nasal segments as desired. Another factor is that their main acoustic features (with the exception of their first formant – see Chapter Three, section 3.4.2) occur within the normal telephone bandwidth of 300 to 3300 Hz, and they are therefore unlike fricatives, which depend on high frequency energy.

Their measureability, however, has proved something of a problem. Their complex spectral structure (including the presence of anti-resonances) and relatively low energy can make it difficult to extract formant features, and Linear Prediction analysis, the favoured tool for Automatic Speaker Recognition, is inherently inaccurate (see Chapter Four, section 4.2.2.1).

2.9.2. Studies of nasality for speaker verification

The number of studies which include nasal segments is rather small. Glenn and Kleiner (1968) first proposed using nasal stops for speaker identification, since they reflect "fixed vocal-tract physiology" and are produced

"with the vocal cavities and the articulators held fixed" (1968: 368). They looked at the alveolar stop [n], using spectrographic analysis of manually located segments. 20 male and 10 female speakers read two word lists in a single session; each list contained 10 words, with the nasal stops [n] and [m] in a variety of positions (initial, intervocalic, preconsonantal and final) and phonetic contexts. Ten tokens of the alveolar nasal in each list were selected, and three spectrographic sections were taken midway through the nasal. These were manually quantised to give 25-band spectra spanning the frequency range from 1000 to 3500 Hz. Each token was then represented by the average of the three spectra. The amplitudes of the spectral bands were normalised, and the whole spectrum was transformed to emphasise its maximum and minimum (equivalent to the major pole and zero respectively). A reference was formed for each speaker by averaging the ten tokens taken from their first word list. The ten tokens from the second word list were used as test items, either singly (ten tokens per speaker), or averaged together. A measure of correlation (the cosine of the angle of separation of the test and reference vectors – see 7.2.1) was used for comparison in an identification experiment. Using all thirty speakers as the population, an identification accuracy of 43% was obtained with ten single test tokens. Forming the test token by averaging single tokens improved the identification accuracy to 62% using average of two, 82% using an average of five and 93% using an average of ten. When the thirty speakers were separated into three groups of ten (two groups of males, one group of females), with comparisons taking place only within a group, an accuracy of 97% was achieved using a 10-token average test token.

Other studies have followed Glenn and Kleiner in adopting a segmental approach, comparing features of nasal stops with features extracted from other, non-nasal segments. Wolf (1972), using filter-bank analysis, and Sambur (1975), using Linear Prediction, examined formant frequencies from the bilabial and alveolar nasals; Paul and co-workers used a variety of spectral features extracted from LPC and Fourier analyses of bilabial, alveolar and velar nasals; while Höfker used spectral, cepstral and Linear Prediction parameters for the same three nasal phonemes in German. Savic and Gupta (1990) used broad-class Hidden Markov Models (see 2.6.3.3) to characterise (unspecified) nasal stops as a broad class for each speaker. Kashyap (1976), in a small-scale study of decision rules rather than the verification parameters themselves, used autoregressive modelling on the bilabial and alveolar stops. Nasal stops have also been used by Wood (1978) and Das and Mohn (1971), but neither gives any indication of the relative usefulness of nasal parameters compared with other segments.

The results of these studies have been rather mixed. Sambur (1975) and Höfker (1977) found that the features extracted from nasal stops were quite useful: in Sambur's study, nasal formant frequencies were among the best five parameters ranked by F-ratio, even though recordings spanned over three-and-a-half years, while Höfker found that the nasal stops gave the lowest identification error rates among the 24 phonemes studied. According to Wolf (1972), Paul and co-workers (1975) and Kashyap (1976), however, nasal stops were no better than other segments such as vowels and voiceless fricatives.

Savic and Gupta (1990), using single session data, obtained the best verification results with voiced segments, which gave an average error rate of 9.3%, compared with 9.75% for fricatives, 17% for nasals and 36% for plosives. However, the success of Hidden Markov Modelling depends on obtaining adequate training data for each model, and it is not clear whether all the models were trained to the same extent.

No serious study has been made of whether the inclusion of nasal segments in utterances for verification using contour or statistical analysis affects the results, though there is a preference for such utterances in much of the verification literature (e.g. Velius 1988). Two studies (McGonegal et al. 1979, Furui 1981a) have compared a wholly oral utterance – the sentence "*We were away a year ago*" with an utterance containing nasal segments – "*I know when my lawyer is due*", but these tell us nothing: McGonegal and co-workers (1979) examined the performance of fundamental frequency and gain, rather than spectral parameters; and while Furui (1981a) looked at the performance of cepstral coefficients, in both cases the difference was confounded with one of *sex*: the males spoke the first sentence, the females the second.

One feature of the studies of nasal stops is that explicit consideration of the anti-resonance or *zero* frequencies has been avoided. Glenn and Kleiner, Höfker and Kashyap simply used the whole spectrum in its various forms. Wolf used the filter-bank outputs which gave the highest F-ratios in a single utterance by each speaker of [m] and [n] in "*I cannot remember it*": those chosen apparently corresponded to the formants at 250 Hz, 2000 Hz and 3000

Hz in [m], and at 1000, 2000 and 3000 Hz in [n]. Savic and Gupta (1990) used three equivalent LPC representations (PARCOR, cepstral and log-area ratio coefficients) plus the gain. Sambur (1975), using Linear Prediction analysis to derive formant frequencies, attempted to overcome the inherent inaccuracy of this approach by using only formants from regions where no spectral zeros were to be expected (the 1000 Hz formant in [n] and the 3rd or 4th formant (between 1700 and 2300 Hz) in [m]. However, there is some evidence that spectral zeros may themselves prove useful: Wolf also found a high F-ratio for the filter-bank outputs "in the region of pole-zero interplay below 1 kHz" (1972: 2049).

2.10. Conclusion

This review has surveyed a wide range of techniques and studies in speaker verification. Studies have been classified according to the acoustic parameters under investigation – gain, source parameters (phonation), filter parameters (articulation) and timing parameters – and according to the methods used – text-dependent contours, text-independent statistics and semi-text-dependent segmentation analysis. The advantages of the segmental approach have been highlighted, and in particular the use of nasal segments. The acoustic characteristics of nasal articulation have received comparatively little attention. Only nine studies can be said to deal with nasal characteristics in any detail - five of them looking at the overall spectrum (Glenn and Kleiner 1968; Su, Li and Fu 1974; Kashyap 1976; Höfker 1977; Wood 1978; Savic and Gupta 1990), and three extracting resonance features (poles or formants) (Wolf 1972; Sambur 1975; Paul *et al.* 1975). This neglect is in spite of the widely

acknowledged potential of nasality for speaker recognition. This potential is borne out by some studies reviewed here (e.g. Glenn and Kleiner 1968, Sambur 1975 and Höfker 1977), but not by others (Kashyap 1976, Savic and Gupta 1990).

The following chapters therefore reconsider the usefulness of nasality for speaker verification, beginning with a review of the anatomy, acoustics and phonetics of nasality in Chapter Three, and paying particular attention to the nature of acoustic variability in the nasal spectrum and the presence of anti-resonance effects.

CHAPTER THREE

NASALITY – A REVIEW

CHAPTER THREE

NASALITY – A REVIEW

3.1. Introduction

It was suggested in the Chapter Two that more work might profitably be done on the use of nasal segments for speaker verification, particularly nasal stops, given their neglect in recent years, and the conflicting findings of the studies which have used them.

The principal argument advanced in favour of the use of nasal segments hinges on the nature and role of the nasal cavities in speech production. Generally speaking, in oral sounds the velum prevents sound energy escaping via the nasal tract and the nostrils, and only the pharynx and oral cavity contribute to the spectrum of the speech wave; in nasal sounds, however, the lowering of the velum allows the nasal tract to resonate, adding its characteristic filtering action to that of the rest of the vocal tract. This filtering action depends on the size, shape and detailed cross-section of the nasal tract (as with any resonator). It is known that the size and structure of the nasal tract varies considerably from speaker to speaker, and it is therefore quite likely that this variation is reflected in the spectra of nasal segments. Perhaps more importantly for speaker verification, the size and shape of the nasal cavity are rela-

tively fixed by comparison with other vocal tract cavities, and cannot be changed at will. Thus the characteristics of the spectrum of nasal segments can be expected to be highly speaker-dependent, and not easily disguised or imitated. It has been pointed out, however, that the response of the nasal tract is also highly dependent upon the speaker's state of health, and that this makes nasal spectra rather less reliable (Sambur 1975).

This chapter reviews some of the literature on nasality and examines the arguments for and against its use in speaker verification. The nature of nasality and the nasal cavities' contribution to it are discussed, and alternatives to the use of nasal stops are considered (3.2). The anatomy and physiology of the nasal tract, and of other structures involved in speech production, are reviewed briefly, and the extent to which their configuration can be varied is discussed, as is the effect of changes in speakers' health (3.3). The remainder of the chapter is devoted to studies of the acoustics of nasality. The acoustic theory of nasality is outlined, and the acoustic properties of the chief exponents of nasality – nasal stops and nasal vowels – are reviewed briefly (3.4). The properties of the vocal tract cavities and their contribution to the nasal spectrum are then considered, paying particular attention to both the extent and the stability of the contribution of the nasal cavities themselves (3.5). The dependence of this contribution on the configuration of the other cavities of the vocal tract, and on the degree of velopharyngeal coupling, is emphasized. Finally, the acoustic consequences of cavity changes are considered, and some data on variability in nasal spectra are reviewed.

Some gaps in our knowledge of the nature and extent of variation in the acoustics of nasality are exposed in this review, and suggestions are made for areas where work might be done. In particular, the choice of a suitable analysis method for characterising nasality and the effect of phonetic context (specifically vowel context) are highlighted for consideration in later chapters.

3.2. Speech production and the production of nasality

3.2.1. Introduction

This section considers the role of the nasal cavities in speech, and in the production of nasality. It is shown that not all manifestations of what is called "nasality" depend on the nasal cavities, and that the conditions for nasal resonance are complex, in that they include but are not limited to velopharyngeal opening, as in nasal stops and nasal vowels. Some differences between nasal stops and nasal vowels are highlighted.

3.2.2. Speech production – a general view

The airstream for speech is provided normally by expired air from the lungs; the vibration of the vocal folds in this airstream provides acoustic energy over a range of frequencies in the form of a series of sharp pulses, one at each instant of glottal closure; this quasi-periodic wave passes up through the pharynx, being modified by the articulators in the mouth – the tongue, lips and teeth – and by the resonance characteristics of the various cavities of the vocal tract* – the pharynx itself, the nasal cavities and the oral cavity. The

* The term *vocal tract* is throughout being used to refer to the whole of the supraglottal cavity structure, including the nasal passages.

modified sound wave is radiated from the lips and/or nostrils, and also by conduction through the chest, the throat wall, the cheeks and the bones of the skull.

3.2.3. Production of nasality

It is generally accepted that nasality is produced when the velum is lowered to allow the sound wave to pass through the nasal tract, as well as or instead of through the oral cavity. Catford (1977: 137) states that

> all sounds produced with the velum lowered (nasal port open) ... are termed *nasal* or *nasalized.*

Similarly, Malmberg (1963: 30) states:

> the movements of the *velum* determine whether a sound will be pronounced with or without *nasal resonance.* If the soft palate closes the passage to the nose by pressing against the posterior wall of the pharynx, we obtain an *oral* articulation. If on the other hand the soft palate leaves this passage free ... we obtain a *nasal* articulation.

The essence of "nasality" in these definitions is that the nasal tract is freely connected to the oral-pharyngeal tract through the velopharyngeal port. The nature of this connection is treated variously (sometimes interchangeably) as *aerodynamic* – expressed in terms of the free passage of air between the oral and nasal tracts – as in Catford (1977: 137), and *acoustic* – expressed in terms of the freedom of the nasal tract to *resonate* – as in Fant's description (1970: 139). Malmberg (1963) talks of both resonance and airflow in his definition.

Resonance and airflow are closely related, but not synonymous. There can be nasal airflow without nasal resonance: Van Riper and Irwin (1958) report

that airflow into the nose through the open velopharyngeal port is a common occurrence in non-nasal speech, and according to Green (1964: 58), "slight escape of air down the nose ... does not necessarily mean vowel nasality". Airflow can also occur even when there is full velopharyngeal closure, owing to movements of the velum altering the volume of the nasopharynx (Lübker and Moll 1965, Smith 1951). Nasal resonance without airflow has been reported by Rosetti (1962: 75), who attributed it to "propagation of the vibration from the mouth into the nasal cavity through the soft structure of the soft palate", according to Cagliari (1978: 136). Laver (1980: 88) also mentions that the absence of airflow in speakers with a "posterior nasal blockage" such as during a head cold does not preclude nasal resonance (though of a different kind), since it is highly likely that sound waves can travel through the mucus or through the tissues of the velum.

It is also true that some types of nasality do not depend on the nasal tract at all. A large body of research, particularly in the area of speech pathology, is devoted to the study of what can be called *nasal voice quality* – that is, an auditory impression of continued nasality throughout a person's speech. The role of the nasal tract in the production of nasal voice has been a matter of much debate. Much of the research in this area comes from speech pathology, whose interest is chiefly in the causes of *excessive* or *inappropriate* nasality. Both can arise from various forms of velopharyngeal incompetence or insufficiency (Hirschberg 1986) leading to genuine nasal coupling, including (but not exclusively) *cleft palate* – a failure of the bones of the facial skeleton

and their associated membranes in the area of the hard palate to fuse properly in the midline during gestation, leaving a space or cleft at any place between the upper lip and the velum, so that complete isolation of the nasal tract is impossible. However, West and co-workers (1957: 199) point out that

> Frequently there comes to the clinic a person whose voice is distinctly "nasal" in quality but whose vowel sounds are made with the nasal port unmistakably shut tight

It is clear from such comments that nasality in this case is an *auditory* quality in which nasal tract resonance need play no part. Examination of the considerable body of literature in this area shows that the label "nasality" is applied to a broad range of phenomena whose common feature is vocal tract resonance of a particular sort (Laver 1980), involving a high degree of damping (Van den Berg 1962), with or without the presence of anti-resonance somewhere in the vocal tract (not necessarily the nasal cavity). West and co-authors (1957: 196-7, cited in Laver 1980: 78) contend that

> The timbre ... usually given the name *nasality* is the result of resonance in a cul-de-sac resonator, a chamber opening off from the passageway through which a sound is resonated and delivered to the outer air.

Various suggestions have been made of possible side-cavities which might impart a nasal quality to a person's speech: Greene (1964: 67), for example, suggested "constriction of the pillars of the fauces" and "muscular contraction in the laryngeal cavity", both creating cavities which could act as cul-de-sac resonators.

West and co-authors (1957) rightly suggest that the term *nasality* is better restricted to resonance involving the nasal tract, and that a separate term such

as *cul-de-sac resonance* should be used for the more general phenomenon, of which nasality is a special type.

3.2.4. Manifestations of nasality

Genuine nasality involving nasal cavity coupling has a wide variety of manifestations and is not restricted to the articulation of nasal stops. Ferguson (1975b), for example, lists nasal stops, nasal vowels, "nasal continuants of various kinds" such as nasalized semivowels, segments with nasal onsets or offsets (prenasalized stops, vowels with nasal offglides and post-nasalized stops), and nasal clicks. According to Crothers (1975: 154), however, "a nasal consonant system is ... basically a stop system". The simple nasal stops are indeed the most common nasal consonant: in Crothers' sample, for example, 104 languages had a simple bilabial nasal, 102 a simple alveolar or dental nasal, and 82 a simple velar nasal, while complex forms appeared in far fewer languages (13 in the case of the prenasalized bilabial stop, for example).

Nasal stops can be made at many places of articulation, from bilabial to uvular. The IPA phonetic symbol chart (International Phonetics Association 1949) provides for seven principal categories: bilabial, labiodental, dental/alveolar, retroflex, palatal, velar and uvular. Individual languages appear to have no more than five places, however (Crothers 1975: 161). Closure locations in the pharynx or at the glottis rule out the participation of the nasal tract and do not occur (Ladefoged 1971: 41).

Voiceless nasal stops occur, in Burmese for example (Ladefoged 1971: 117), but are relatively rare. They are more likely to be bilabial than any other

place of articulation (Maddieson 1984: 69); Catford (1977: 138-9) suggests that they might more appropriately be regarded as "nareal fricatives", since their acoustic energy comes almost exclusively from turbulence created around the nostrils. The influence of the nasal cavities on their spectrum will therefore be minimal. The fact remains that they are accompanied by oral constrictions at different places (bilabial, alveolar and velar in the case of Burmese) and they probably are best regarded as stops. Partial devoicing of nasal stops in voiceless contexts is more common: Gimson points out that British English (RP) nasal stops may be partially devoiced after [s], while syllabic velar [ng], occurring after [k], may be partially devoiced too (1970). Nasal vowels occur as phonemes, contrasting with corresponding oral vowels, in many languages such as Yoruba (Ladefoged 1971), French and Bengali (Ferguson 1975b), but they are reported to be much less frequent in their occurrence than their oral equivalents (Ferguson 1975). Some languages appear to have more than one degree of distinctive nasality: Chinantec, for example, is reported to have contrasts between oral, lightly nasalized and heavily nasalized vowels (Merrifield 1963, cited by Ladefoged 1971). Many languages without distinctive nasality in vowels nevertheless have nasal vowel realizations, usually attributable to *coarticulation* effects. Thus vowels adjacent to nasal stops in American English are normally nasalized (see 3.2.7). Nasalization appears to have an effect on the perceived height of the vowel (Wright 1975, Beddor 1986, Krakow et al. 1987), though whether this is due to a change in articulation (as a result of the mechanical linkage between the tongue and velum, for example) or has a purely acoustic explanation is a matter for debate.

3.2.5. The universal nature of nasality

Nasality has a segmental function in most if not all known languages, principally in the form of nasal stops. Ferguson (1963, cited by Maddieson 1984: 61) suggested, as a possible phonological universal, that "every language has at least one primary nasal consonant in its inventory" – that is, a phoneme whose most characteristic allophone is a nasal stop. Several languages appear to have no primary nasal stop phonemes, but many still possess nasal segments of some description: Maddieson (1984: 61), for example, reports that ten of the 317 languages in the UPSID sample have no *primary* nasal consonants, but that only four of these ten have "no phonemic nasal or nasalized segments of any kind"; the other six have pre-nasalized stops (see below) or nasalized vowels. Abercrombie (1967) gives Wichita as an example of a language with no nasal consonants, but does not say whether it has any other nasal segments instead.

These descriptions refer to phoneme inventories, as Maddieson is careful to point out, and may not be reliable indicators of the status of nasality in a language. Nasal *segments* do occur in at least some of the languages described as having no nasal consonant phonemes (Maddieson 1984), while languages recorded as having nasal phonemes may actually be extremely restrictive in how they are used. Ferguson (1975: 176) cites Puget Sound Salish as "a language in which nasality hardly seems to function at all", despite the existence of both bilabial and alveolar nasal stops, explaining:

> In ordinary adult speech only the word for 'little' [mi?ma?d] or [mi?ma?n] has nasal segments. In baby talk register nasality occurs in that voiced stops are

replaced by nasal stops in certain diminutives. Also in the religious register and in story-telling nasality appears; for example, certain stock characters speak with nasal instead of oral voiced stops.

3.2.6. Denasality

Phonologically nasal segments are not always pronounced with nasal resonance by all speakers. The absence or weakness of audible nasal resonance where it would normally be required is termed *denasality*, and may stem from a variety of causes. Very rarely, however, does it preclude nasal cavity resonance altogether.

Pathological causes of auditory denasality (or *hyponasality* as it is often referred to in the speech pathology literature) may include nasal cavity blockages caused by cattarh or inflammation; a deflected nasal septum; growths such as polyps or adenoidal swellings which obstruct the nasal tract; or an excessively long uvula (Cagliari 1978). As Laver (1980:89) observes, some of these conditions (e.g. that of a head-cold) give rise not to the absence of nasal resonance but to "a special, very highly damped kind of nasality". Presumably in such cases the cavity behind the nasal tract obstruction functions as a cul-de-sac resonator, with or without the participation of the passages beyond the obstruction. In the case of adenoids (pathological enlargement of the pharyngeal tonsil in the nasopharynx: cf. 3.3), the nasal port may be completely obstructed (Kaplan 1971) – though at exactly what level is not quite clear – and presumably a genuine lack of nasal coupling can occur.

A speaker may choose to speak with the velopharyngeal port closed, in which case there will indeed be a lack of nasal resonance in phonologically

nasal segments, other than resonance resulting from transmission through the velum and hard palate. Such speech would be perceived as highly deviant, however.

Denasality is not used linguistically for either segmental or suprasegmental purposes, but it may be an personal idiosyncrasy, or function to signal membership of a speech community, as in the example of Liverpool speech (Abercrombie 1967, Knowles 1978). The mechanism of such denasality is not clear: Abercrombie suggests that an adenoidal (denasal) voice quality can be simulated with "continuing velic closure, together with velarization" (1967: 95), but a study of voice quality among Liverpool speakers by Knowles (1978) mentions only a raising and retraction of the tongue body, constriction of the faucal pillars, raised larynx, constricted pharynx and a close jaw position. Laver (1980: 92) mentions that, paralinguistically, denasality "is sometimes heard as a signal of incipient laughter", presumably as a result of the reflex control of the velum over-riding its linguistic use.

3.2.7. Suprasegmental uses of nasality

The literature has tended to concentrate on segmental occurrences of nasality but nasality is a suprasegmental feature too.

Nasality is prone to spread into neighbouring segments. Vowel nasalization is very commonly reported in the environment of nasal stops, especially in languages with no phonologically contrastive nasalized vowels such as American English. It has been shown, for example, that the velum begins to lower well before the formation of the oral constriction for the nasal stop (Moll and

Daniloff 1971). Ohala (1971, cited by Bell-Berti 1980) reports that this effect is greater in vowels preceding nasals than in vowels following nasals. Nasalization of a vowel before a nasal stop can in some instances take the place of the nasal stop. Fant (1973: 154-5) gives the example of a word such as *wing*, which in some dialects of American English has a heavily nasalized vowel but no actual stop constriction.

Ladefoged (1974: 33) points out that nasality is frequently a property more of whole syllables than of individual segments, and he seems to suggest, in fact, that some so-called nasal segments – semi-vowels, fricatives and laterals, for example – are nasal simply because the syllable in which they occur is nasal.

Nasality is also seen to operate in larger speech units than the syllable. In Sundanese, for example, it functions as a marker of verb forms: according to Laver (1980: 3-4),

> once initiated by a nasal consonant in any position in the syllable, nasality in Sundanese runs forward through all syllable boundaries until checked by a word boundary or a supraglottal consonant (Robins 1953, 1957).

Suprasegmental nasality also has a *paralinguistic* function in some communities, expressing subordinacy in Cayuvava, where vowel nasalization is used by a speaker of lower social rank addressing one of higher social rank (Laver 1980), and as a marker of politeness in honorifics in Bengali (Ferguson 1975). Ferguson (1975: 176) notes that in Puget Sound Salish it is restricted in its occurrence to particular registers (such as baby talk) and diminutives. In language in general, nasality often features in hesitation markers. Laver also points out that it serves as a marker of group identity for some speech com-

munities, characterizing the speech of some dialects of American English and British English.

3.2.8. The function of the velum in nasality

The velum, which controls the coupling between the nasal, oral and pharyngeal cavities, does not actually function in the on-off manner implied by much of the preceding description. Several researchers have shown that the position of the velum varies with both the content and the context of a segment, though their methods do not allow any statements to be made about velopharyngeal aperture. Künzel (1979), for example, investigated velic height in non-nasal segments by indirect observation, tracking the amount of light reflected from the superior surface by means of a flexible probe with a photoelectric transducer and light source, inserted into the nasal cavities. Velic height was least for vowels, and greatest for plosives, with the approximant [l] in between. Voiceless stops showed a higher velic position than voiced stops; and velic height seemed to increase with a more back articulation, with labial stops (when orally released) having the lowest position and velar stops the highest. Such differences might be explained by changes in intra-oral pressure, which has been shown to be higher in voiceless stops than in voiced stops and higher in back articulations than in front articulations (Minifie et al. 1973: 265-6); if so, they would presumably *not* be present in nasal stops, in which there is no build-up of intra-oral pressure.

Moll (1962), Lübker (1968) and Fritzell (1969) have found that velic height is greater for high vowels than for low vowels, confirming the reports of

Czermak (1857, 1858, 1869) and Passavant (1863) (both cited in Bell-Berti 1980) that velic height increases through the vowel series /aa,e,o,uu,ii/.

Cagliari (1978: 159), drawing on reports from a variety of sources (e.g Condax et al. 1976, Kaplan 1960, Calnan 1955, MacNeilage 1972, Bell-Berti and Hirose 1975, Harrington 1944), proposes that a scale of velic height in normal speech can be established, as follows:

highest position of velum	blowing
	voiceless stops
	voiced stops
	voiceless fricatives
	voiced fricatives
	oral close vowels
	oral open vowels
	nasalized close vowels
	nasalized open vowels
	nasals (i.e. nasal stops)
lowest position of velum	respiratory position

While there are statements in the literature that nasalized vowels have a lower position than oral vowels, and open vowels have a lower position than close vowels, it is not clear that the relationship between oral open vowels and nasalized close vowels is exactly as stated. What is clear from this scale, though, is that nasalized and nasal segments have the lowest velic positions other than that for breathing itself.

Cagliari proposes (1978: 164) that the behaviour of the velum even in nasalized speech can be accounted for by treating this scale as a *neutral velic scale* describing the relationship of various segments. Thus the addition of a long-term setting of nasalization to speech does not result in a continuous opening of the velopharyngeal port, but a shifting downwards of the neutral velic

scale such that all or most segments in the scale have greater velopharyngeal opening, with their general order on the scale preserved. Similarly, a setting of *denasality* constitutes an *upward* shift, with a greater degree of velopharyngeal closure on all segments, but with the position of the velum still varying according to the order determined by the scale. This suggestion appears to be borne out by measurements of nasal airflow in the normal, nasalized and denasalized speech of a single speaker (Cagliari 1978: 262-265), though there is some disruption to the ordering of the segments on the scale. The validity of using airflow as an indicator of velic height and velopharyngeal opening is open to question, however, since nasal airflow is, according to Lübker and Moll (1965: 271), "dependent not only upon the amount of velopharyngeal opening, but also upon the amount of oral constriction". It is possible that the production of nasalization affects the degree of oral constriction in a given segment, given the need, for example in fricatives, to maintain a turbulent flow of air through the constriction. There is also evidence that "nasal" speakers retract and raise their tongue more than normal speakers (Hixon 1949, cited by Laver 1980), though whether this is a normal feature of nasalization in its segmental function is not clear.

Variation in velic position (and velopharyngeal aperture) apparently occurs even in phonologically nasal segments. Kiritani and co-workers (1980), for example, in an X-ray microbeam study of nasal stops in Japanese, found differences in velic elevation depending on the position of the stop within the syllable, while Bell-Berti and co-workers (1979a) found a vowel height effect,

with the velum being higher for both oral and nasal stops in the context of close vowels.

3.2.9. Velopharyngeal opening and nasal resonance

It is evident from the preceding description that velopharyngeal opening alone is not enough to guarantee the production of audible nasal resonance. The degree of opening required for the production of nasal resonance has been examined in several studies. Perceptual studies using synthetic speech (e.g. House and Stevens 1956) have shown that a greater degree of velic opening is required to induce perception of nasality in open (low) vowels than in close (high) vowels; nasal stops too need a greater degree of velopharyngeal opening when they precede low vowels if they are to be perceived as nasal, rather than oral, stops (Abramson et al. 1981). Unfortunately, much of the work – especially in the speech pathology literature – has attempted to discover the threshold of velopharyngeal opening which introduces *inappropriate* or *abnormal* degrees of nasality into oral speech. Passavant (1863), for example, inserted rubber tubes of different diameters into the nasopharynx of speakers to maintain a velopharyngeal opening of known aperture when the velum was raised; according to Fritzell (1969: 8), an opening with an area of 12.6 mm sq. "did not appreciably influence speech", while an area of 28.3 mm sq. "gave most of the consonants a nasal character, but the vowels were still not influenced". Kaltenborn (1948) measured the extent of both velopharyngeal opening and oropharyngeal opening from X-ray pictures in speakers who were perceived as "nasal", as well as in normal, non-nasal speakers. According to Laver (1980: 79),

> the typical size of the opening to the nose (presumably on the front to back diameter) for non-nasal speakers was 1 mm, and for the opening to the mouth 11 mm; the measurements for speakers judged as nasal were 8.8 mm for the opening to the nasal cavity, and 3.1 mm for the opening to the mouth.

Kaltenborn concluded from his observations that

> [Abnormal] Nasality is caused by having too wide an opening into the nasopharynx in comparison with the opening into the oral cavity

(quoted by Van Riper and Irwin 1958: 241).

On the basis of studies of both normal and abnormal nasality, several writers (Kaltenborn 1948; Van Riper and Irwin 1958; Laver 1980) have suggested that a crucial factor in determining the presence of nasal resonance is not the size of the velopharyngeal aperture itself, but the relationship (or ratio) between this aperture and the aperture from the pharynx into the *oral* cavity.

More generally, it has been proposed that speakers attempt to maintain "a characteristic balance or ratio between oral and nasal resonance" (McDonald and Baker 1951: 11), which Abramson and co-authors (1981: 330) interpret as suggesting that "a suitable ratio of acoustic *impedances* of the nasal tract and the oral tract is necessary".

The exact nature of this ratio has been the subject of some speculation. According to Laver (1980: 83):

> the side chamber will generate audible nasality only when the entry to the side chamber has an area approximately equal to or greater than that of the entry to the other cavity.

Van Reenen (1982) has attempted to formalise the relationship between the degree of nasal coupling and the coupling to the oral cavity, suggesting that it is the ratio of the cross-sectional area of the velopharyngeal port to the sum of

the areas of the velopharyngeal port and the oral constriction which determines whether a sound is nasal or not. He based his work on estimates of these parameters made from a large number of X-ray studies of nasal stop and nasal vowel production in various phonetic contexts. The ratio,

$$N\% = N/(MC + N) \cdot 100 \tag{3.1}$$

where N is the area of the nose-coupling in mm^2 and MC is the area of the mouth constriction, is evaluated for a variety of nasal segments in various phonetic contexts. He found that nasal vowels are generally characterised by

> a change in N% from (almost) zero to about 75 and an increase in the amount of nose coupling N

(1982: 135). Nasal stops, however, were always produced with N% of 100, since the area of the mouth constriction was always zero. Van Reenen observes that, because changes in velopharyngeal coupling in nasal stops cannot affect N%,

> there is no specific amount of nose coupling characteristic of nasal consonants, and the velum may move quite freely and adopt a great variety of shapes and positions

(1982: 131), so long as some nose coupling remains.

According to his formula, while the ratio for nasal vowels will vary, depending on changes to the oral and nasal coupling areas, for nasal stops it will always be 100%, since the area of the mouth constriction is always 0. Thus in this sense, the degree of nasal coupling will have no effect on the perceived nasality of the stop.

3.2.10. Summary

The selection of nasal stops as the segments in which to look for the spectral characteristics of the nasal cavities has not previously been discussed. The choice has been made rather on practical grounds: the languages studied (American English and German, for example) had no other nasal segments. Consideration of aspects of the production of nasality, however, bears out this choice. While there are many manifestations of nasality in speech, and widespread nasal coupling even in "non-nasal" speech, resonance of the nasal cavities is an essential feature only of phonologically nasal segments such as stops and vowels. Nasal stops appear to be superior to nasal vowels: they are practically universal in language, they have the greatest degree of velopharyngeal opening (Cagliari 1978) – though this can still vary with context – and since they are stops, by definition the ratio of nasal coupling to oral coupling during their production (Van Reenen 1982) is at its greatest and most stable.

3.3. Anatomical and physiological variation in the nasal tract and other vocal tract cavities

This section considers the anatomy and physiology of the vocal tract, its potential for movement and some causes of variation in its shape and size. The discussion is not restricted to the nasal cavities alone, since it is desired to explore how invariant they are by comparison with the other cavities. Much of the anatomical detail is drawn from works by Zemlin (1968), Kaplan (1971) and Romanes (1986), supplemented by material from Laver (1980) and McMinn and

Hutchings (1988).

The supraglottal vocal tract consists of three main cavities: the pharynx, the oral cavity and the nasal cavity. There are also numerous minor cavities (the sinuses) associated with the nasal cavity, and several small recesses within the pharynx formed by various membranes. All are capable of being varied in size and shape to some extent, but the nasal cavities are indeed the least capable of voluntary movement.

The *pharynx* is a very active and mobile cavity, and its length, volume and cross-section are all variable. A large part of its structure is made up of muscles, which are arranged in two layers in the lateral and posterior walls. Only in the uppermost part of the pharynx wall, between the levator palati muscle and the base of the skull, do the muscle fibres give way to fibrous tissue, the *pharyngeal aponeurosis*. The action of the pharyngeal muscles changes both the diameter of the tube and its elevation. The fact that the front wall is largely formed by other structures means that its shape is greatly affected by their movements too. Tongue movements possibly have the greatest effect, not just because neighbouring structures (such as the epiglottis, or the tongue dorsum) are moved into the space of the pharynx, but because the muscles involved also have a role in maintaining the conformation of the pharynx (e.g. the middle pharyngeal constrictor, which can raise the hyoid to aid tongue-fronting/tongue-raising: Laver 1980: 26; Van Riper and Irwin 1958: 366). Raising and lowering of the *larynx* by its extrinsic musculature also affects the length and diameter of the laryngopharynx (Kaplan 1971: 224).

The *nasopharynx* or *epipharynx*, that part of the pharynx to the rear of the nasal cavity, is perhaps the least mobile part. It is the widest part of the pharynx, about 4 cm wide and 2 cm from front to back (Zemlin 1968: 304). Its upper boundary is the base of the skull, and therefore fixed, but its lower boundary is formed by the upper surface of the velum and is therefore variable depending on the degree of velopharyngeal closure. It opens anteriorly into the nasal cavities through the the *posterior nares* or *choanae*. When the velum is lowered, as in the production of nasality, the nasopharynx is continuous with the oropharynx; when the velum is raised to meet the posterior wall of the oropharynx, the nasopharynx is isolated. The nasopharynx is therefore in *permanent* communication with the nasal cavities (through the choanae), but in *variable* communication with the rest of the pharynx depending on the position of the velum. The lateral walls of the nasopharynx also contain muscle fibres, and are capable of a certain degree of movement medially. This movement has been observed by many researchers (e.g. Harrington 1944, Calnan 1955, Dickson and Dickson 1972, Zagzebski 1975), and appears to be consistently present during velopharyngeal closure. On its posterior wall there is a variable amount of defensive lymph tissue forming the *pharyngeal tonsil*.

The shape and volume of the *oral cavity* are highly variable. Though part of its structure is skeletal – the hard palate and the teeth, for example – the movements of the jaw, tongue, lips and velum all affect its shape. The tongue is especially mobile, being free to move anteriorly, laterally and superiorly, and it has a complex muscular structure which allows it to assume a great variety

of configurations, affecting the shape not just of the oral cavity but of the pharynx too. Its attachments posteriorly and inferiorly to the hyoid bone, the epiglottis, the velum, the mandible and the pharynx mean that its movements generally have consequences elsewhere in the vocal tract.

The *nasal cavities* lie anterior to the nasopharynx, and their framework is entirely skeletal, being formed by the bones of the skull and the facial skeleton. The picture given of the nasal cavities in some phonetics textbooks, showing a single large open cavity above the hard palate (e.g. O'Connor 1973: 30, Fig. 6), is misleading. They are rather *two* high, narrow cavities divided throughout by a median bony septum. They are about 8 cm long, 5 cm high and roughly 10-12 mm wide at the floor of the cavity, narrowing at the top to only 1 or 2 mm for most of its length. Their dimensions and detailed structure are known to vary widely across individuals (e.g. Lindqvist and Sundberg 1976). Typically, the septum deviates to one side, so that the two nasal cavities are rarely symmetrical (Romanes 1986: 149).

The twin cavities have a complex internal structure. Two bony structures – the lateral processes of the ethmoid bone – project downwards from the roof, one into each nasal cavity. Each lateral process consists of two complex turbinated ("scroll-like") projections of bone, the superior and middle nasal *conchae*. These cover the lateral wall of the cavity, and, together with the *inferior nasal concha*, an independent bone of similar structure attached to the lower lateral wall, they divide each cavity into three connecting channels, the *superior, medial* and *inferior meati*. In some persons, a small fourth concha (the

supreme) is present, above the superior concha (McMinn and Hutchings 1988: 45).

The complexity of the nasal cavities is increased by their communication with several air-spaces or *sinuses* within or between the bones of the skull. The *maxillary* sinus pair is the largest; they are hollows inside the cheek-bones, immediately adjacent to the nasal cavity proper, with one or two openings in the lateral nasal wall, between the inferior and middle nasal conchae. The opening of the sinus allows fluid to drain out of the sinus into the nasal cavity, but because it is located rather high up in the sinus wall, a large amount of fluid can collect within when the head is upright (Romanes 1986: 156), reducing the volume of air in the sinus. The *ethmoidal* sinuses lie between the upper part of the nasal cavities and the orbit (eye-socket), and consist of many smaller air-cells (Kaplan 1971: 223) divided into three groups: the anterior and middle cells open into the nasal cavity on the lateral wall, below the middle nasal concha (into the middle meatus), while the posterior cells open above the middle concha (into the superior meatus). The *sphenoidal* sinuses are paired, though often highly asymmetrical, chambers located posterior to the nasal cavity proper, above the nasopharynx. They drain into the nasal cavity by a small hole, typically covered with a flap of mucous membrane (Romanes 1986), above the superior nasal concha. The paired *frontal* sinuses are hollows in the frontal bone of the cranium; they are highly asymmetrical, and may extend laterally above the orbit as far as its outer edge. They communicate with the nasal cavity via a hole in the anterior part of the lateral

nasal wall, below the middle nasal concha and forward of the opening to the maxillary sinus.

The external framework of the nose is the only part forward of the nasopharynx whose shape can be varied by muscular action, and even this movement is limited to dilation and constriction of the nostrils (Zemlin 1968: 264). It consists of plates of cartilage, fatty tissue and muscles, built on to the bones of the facial skeleton and the septal cartilage. It encloses the *vestibule* of the nose (the portion of the nasal cavities anterior to the nasal concha), and ends below at the nostrils.

While the overall dimensions of the nasal cavities are fixed for a speaker, their volume and cross-sectional area are highly variable owing to the activity of the mucous membranes lining the nasal conchae and the lateral walls. These membranes help to filter, warm and humidify the incoming air, and are generously supplied with blood vessels and mucus-secreting glands, under the control of the autonomic nervous system. The secretion of mucus helps to humidify the air and aids the filtering action of the ciliated epithelium; the cilia sweep this mucus, together with any foreign bodies trapped by it, backwards into the nasopharynx, but physiological states which prevent the cilia from operating, such as viral infections (the "common cold") or asthma, may lead to a build-up of mucus and therefore to a narrowing of the nasal meati. A similar narrowing is effected by the enlargement of the mucous membrane with blood, a reflex action which occurs very rapidly in response to a cooling of the air coming into the nasal cavities. The flow of blood is also affected by emo-

tional factors (sudden fright causes vasoconstriction, for example, while anxiety causes vasodilation), and by posture, exercise, hormonal activity and changes in humidity. Certain cells of the nasal lining (the "goblet" cells) can also swell in response to irritation or infection.

Even when atmospheric conditions and emotional state are relatively constant, however, all speakers show an alternating pattern of vasoconstriction and vasodilation in the mucosal lining of each cavity, lasting an average of two and a half hours – the *nasal cycle* (Stoksted and Khan 1976). This alternation allows the nasal lining to maintain an effective balance between its functions of warming and filtering inspired air, since reducing the blood flow allows mucous secretions to maintain the filtering action of the ciliated cells. The pattern of alternation is regulated so that vasodilation in one nasal cavity (with accompanying constriction of the nasal passages themselves) is matched by vasoconstriction (with increased patency of the passages) in the other. In this way, the overall resistance of the airways is held roughly constant (Stoksted and Khan 1976). The length of the cycle and the size of the changes vary with atmospheric conditions (becoming shorter and more pronounced with unfavourable temperature and humidity), with posture, and with the individual: Stoksted and Khan (1976: 519) talk of

> an individually characteristic nasal cycle, occupying from about 30 min to about 5h from one corresponding phase to another ... present in about 80 per cent of normal individuals while in an upright position

Activity of the cycle also varies with age, being greatest in adolescents but decreasing in older people. Disturbances to the nasal passages, such as during

acute sinusitis or with a marked deflection of the nasal septum, also cause irregularity in the length of the cycle and a magnification of the extent of the response.

All parts of the vocal tract show some variation, then. The pharynx and oral cavity are the most variable, being composed largely of muscle or elastic fibres, and are greatly affected by the movements of the tongue. The nasopharynx is rather more limited in its movements, while the nasal cavities themselves, having no muscles, are indeed immovable. They are nevertheless capable of considerable change over relatively short periods of time. This change is largely under reflex or hormonal, rather than voluntary, control, however, unlike the changes affecting the oral cavity and pharynx.

3.4. The acoustics of nasality: nasal stops and nasal vowels

3.4.1. Acoustic theory of nasality

In non-nasal articulations, the sound source (either the periodic glottal wave or aperiodic, turbulent airflow) is filtered by the resonating properties of the combined pharyngeal and oral cavities, acting as a single acoustic tube. This filtering action amounts to a reinforcement of certain frequencies of vibration which coincide with the natural resonant frequencies of the vocal tract tube, and the relative diminution of others which do not. The frequency locations which are reinforced in the spectrum of the resulting sound are known as *formants*. The filtering action of the vocal tract can be described by the locations and bandwidths of these formants (as is customary for vowel

classification), or, in engineering terms, by the distribution of the *poles* and *zeros* of the *transfer function* of the vocal tract: that is, a mathematical expression of its frequency response. In the case of an unbranched resonator, as with most oral articulations (laterals being one exception), this transfer function contains only poles, corresponding to the locations of the vocal tract resonances.

In *nasal* articulations, the opening of the velopharyngeal port adds another filter system with its own resonant frequencies: the nasal tract. This coupling of the nasal tract to the oral-pharyngeal tract causes a radical change in the acoustic properties of the vocal tract as a whole, since there are now two paths for the acoustic energy to take: the nasal tract and the oral cavity.

In acoustic theory, this branching of a resonator system introduces zeros (numerator coefficients) into the transfer function: that is, anti-resonances are introduced into the frequency response of the system. Anti-resonances occur because one branch of the system acts as a "shunt", trapping energy at its natural resonant frequencies instead of allowing it to be transmitted; thus the output from the main branch contains dips at these frequencies, as well as peaks at the frequencies reinforced by the system as a whole. The side-branch also contributes additional poles, and causes a shift in the frequencies of the existing poles of the unbranched resonator system (Fant 1970: 145).

Either the nasal tract or the oral cavity can function as the side-branch, with different acoustic consequences. According to Laver (1980: 83), what determines this is the cross-sectional area of each branch's exit and entrance: whichever cavity has the smaller exit becomes the side-branch, provided that

the exit from that cavity is itself smaller than the entrance. In nasal stops, the side-branch is formed by the closed oral cavity, and the combined nasal-pharyngeal tract acts as the main pathway for sound; velar and uvular nasal stops are a special case, since the oral stricture is at the back of the mouth and the side-branch is reduced to a minimum: in effect, the nasal-pharyngeal tract forms an unbranched resonator system in such stops (Fant 1970: 139). In nasal vowels, it is generally the nasal tract which forms the side-branch, though this is not the case for all vowels (Harrington 1988) since it is possible for the cross-sectional area in the oral cavity to be reduced either at the entrance to the cavity or at the lips sufficiently for the nasal tract to act as the main acoustic pathway, depending on the degree of velopharyngeal coupling.

Both nasal stops and nasal vowels share the spectral complexity caused by this splitting of the acoustic tube into a main branch and a side branch, then; but since the cavities function differently in each case, their acoustic characteristics are very different.

3.4.2. Acoustic characteristics of nasal stops

A prominent low-frequency resonance is the main feature mentioned in the literature (House 1957, Hattori et al. 1958, Fant 1970, Fujimura 1962, 1963; Glass and Zue 1986), and sometimes almost the only feature visible on spectrograms. The frequency of this resonance lies between 200 and 400 Hz. The energy in nasals appears to be concentrated in this region: Fant (1970) notes the "dominating intensity level" of the first formant, while other writers (e.g. Hattori et al. 1958) speak of a "reinforcement" of intensity in this area.

According to Fujimura (1963), the apparent low-frequency boost is more accurately viewed as the result of a second commonly-reported feature – the suppression of energy in the middle frequency range. Glass and Zue (1986: 2768) noted "an abrupt decrease in amplitude in the frequencies immediately above the low-frequency resonance" in cepstrally-smoothed short-time spectra. The presence of a prominent *anti-resonance* has also been noted (House 1957, Hattori et al. 1958, Fant 1970). This may contribute to the separation of the low-frequency resonance from the upper formants.

The upper formants themselves are not always visible on spectrograms (Fant 1973: 27), but useful information has come from Fujimura's analysis-by-synthesis study (1962, 1963). He notes that the formants are less widely spaced (an average of 800 Hz apart) than in vowels, and that they show greater damping (that is, higher bandwidth). One consequence of these properties is a general lack of detail and a low level at higher frequencies. Fujimura notes

> an even distribution of the sound energy in the middle-frequency range between, say, 800 cps and 2300 cps. There is neither a prominent energy concentration nor an appreciably wide and deep spectral 'valley' in any portion of this frequency range

(1962: 1874).

Nasal stops are often reported to be weaker in energy overall than adjacent vowels (e.g. Glass and Zue 1986). House (1957) found that the output of his vocal tract model was indeed lower for nasal stops (between 6.5 dB and 8 dB down on [i], his lowest-energy vowel). Nasal stops are generally stronger than voiced oral stops, however (Fujimura and Lindqvist 1971, Glass and Zue 1986), as might be expected from the fact that they have an opening to the

outside air throughout their production, albeit through the nose.

The low frequency energy peak in stops is usually described as a single formant. Fant (1970), for example, lists the typical formant peaks of nasal stops as occurring at 250, 1000, 2000, 3000 and 4000 Hz approximately. According to Fujimura and Lindqvist (1971), however, this first peak is actually more complex, and their sweep tone study shows it to contain more than one transfer function pole. Further evidence for the complexity of this peak comes from Abramson et al. (1981), who report that the synthetic alveolar nasal stop [n] used in their experiments on the perception of nasality had two low frequency poles – one at 200 Hz with a bandwidth of 80 Hz, and a second at 500 Hz with a bandwidth of 200 Hz – which merged into a single peak at 291 Hz (bandwidth 200 Hz) if the resolution of the LPC analysis was reduced.

According to Kyttä and Hurme (1982: 206) the acoustic characteristics of nasal stops include the presence of an oral formant structure, continuous with that of vowels and visible on spectrograms, "especially if the nasal is between two vowels". Fant (1970) too mentions a "residue" of the oral formant pattern during nasal stops; these formants are contributed by the oral cavity, but are "severely weakened owing to the fairly close proximity of zeros" (1970: 147).

Differences between stops made at different places of articulation appear to be minimal in spectrograms, other than in the nature of the transitions in neighbouring vowels (e.g. Liberman et al. 1954). Detailed analysis even simply of spectral sections does show some differences, however, principally in the location of the major *anti-formant* (Fujimura 1962). Normative data on the for-

mant and anti-formant frequencies are hard to come by, however, partly because of great variation between speakers.

In general, the bilabial nasal shows a lower first formant than places further back in the mouth, usually below 300 Hz (Fujimura 1962: 1870; House 1957; Fant 1970); the first formant of the alveolar nasal is similar in frequency (Fant 1970) or a little higher (Fujimura 1962). The velar nasal almost always has the highest first formant (300 Hz according to House 1957, Fant 1970; 350 Hz in Fujimura 1962).

The number and locations of other resonances vary greatly from study to study, and may reflect the sensitivity of the analysis method used. Tarnoczy (1948), for example, reported peaks at 800 Hz and 2600 Hz for [m], 2300 and 3200 Hz for [n] and 600 and 2200 Hz for [ng]. Modelling studies have provided more detail, but again show rather variable results. For the bilabial nasal, Fujimura reported formant peaks (above F1) at between 800 and 1050 Hz, between 1000 and 1500 Hz and at 1900 Hz; Fant (1970) found peaks at 950, 2200 and 2800 Hz. For the alveolar nasal, Fujimura reported peaks at 1000, 1400, 2300 and 2600-2700 Hz, which match Fant's (1970) figures of 1100, 1400, 2500 and 3000 Hz. For the velar nasal there is good agreement between Fujimura and Fant, despite the differences in their methods: Fujimura lists peaks at 1050, 1900 and 2750 Hz, and Fant at 1000, 2200 and 2900 Hz.

The location of the major anti-resonance appears to be the major distinguishing factor among places of articulation. Hattori et al. (1958), for example, observed energy "valleys" in narrow-band spectral sections over a wide

band of frequencies – from 500 to 1000 Hz in [m], and from 400 or 500 Hz up to as much as 1500 Hz in [n]; for [ng], they report "no dominant absorption", but instead a "dull and complex" spectrum (1958: 272). Fujimura reports an anti-resonance in [m], varying between 750 and 1250 Hz according to vowel context, and in [n] at around 1600 Hz, but no anti-resonance below 3100 Hz (the limit of his analysis) for [ng]. Fant reports two anti-resonances for [m] (at 800 and 3500 Hz) and for [n] (at 1800 and 5600 Hz), but again none for [ng]. House (1957) gives anti-resonance frequencies of 1000 Hz for [m], 3300 Hz for [n] and greater than 5000 Hz for [ng]. It is clear from these and from spectrographic studies of stops at other places of articulation (e.g. Recasens 1983, Cagliari 1978) that the frequency of the major anti-resonance generally rises as the place of articulation moves further back in the mouth.

Fujimura (1962) observed considerable variability in the spectra of the bilabial, alveolar and velar stops, even within a single intervocalic token, and it is clear from his results – and from the way they are presented – that a definitive description of the formant frequencies of the nasal stops is quite difficult. In the case of the velar nasal [ng], spectral matches for three speakers gave a wide range of frequencies: 250 to 400 Hz for F1, 950 to 1150 Hz for F2, 1700 to 2200 Hz for F3 and 2300 to 3000 Hz for F4. Fujimura and Lindqvist (1971) observe that "the transfer characteristics of nasal consonants ... varied greatly from subject to subject", and do not give specific details of any one articulation.

Some authors have found a rather more complex spectrum for nasal stops. Fujimura and Lindqvist (1964a), for example, report on a sweep-tone analysis of various prolonged consonant articulations, including the velar nasal stop [ng] (which they term *palatal*). The pole-zero fit to the response of the vocal tract at low frequencies is complex, with poles (for one speaker) at 300 and 450 Hz and a zero at 380 Hz. The sine-wave response also shows two peaks between 1000 and 1500 Hz, but these are merged into a single peak at around 1300 Hz by the pole-zero fit. A third major peak can be seen at 2000 Hz, and a fourth at around 2300 Hz.

Nord's (1976b) study of Swedish nasal stops using a pole-zero fit to FFT spectra shows a similar feature. Poles can be seen at 250, 450, 1000, 1400, 2000 and 2500 Hz for [m], and at 250, 450, 1150, 1500, 2000 and 2750 Hz for [n̼] (estimated from his Figures), with zeros at 300, 800 and 1700 Hz for [m] and at 400, 1000 and 2200 Hz for [n̼]. Again, the complexity of apparently simple formant peaks can be seen, with the 200-400 Hz first formant breaking down into two poles and a zero in the pole-zero fit.

3.4.3. Acoustic properties of nasalized vowels

It has been observed (Delattre 1969a, 1969b) that, acoustically, nasal stops and nasalized vowels have little or nothing in common. Nasalized vowels are typically described in terms of changes to the corresponding oral vowel spectrum, which is regarded as having nasality superimposed on it (e.g. Joos 1948, Fant 1973: 27). Their acoustic description is complicated by the fact that the effects of vowel nasalization vary with the speaker, the vowel and the degree of

nasal coupling (Fant 1970). General characteristics which have been noted in the literature include a reduction in the intensity of the first formant of the vowel (e.g. Schwartz 1968; Delattre 1954, 1969a,b; Smith 1951; House and Stevens 1956; Glass and Zue 1985). and increased bandwidth of the vowel formants (House and Stevens 1956: 221; Glass and Zue 1985). Additional resonances and anti-resonances are introduced into the spectrum, principally in the area of the first vowel formant (Smith 1951, Hattori et al. 1958, Glass and Zue 1985); and there is a shift in the frequency of the oral vowel formants (Fant 1970, Fujimura and Lindqvist 1971).

Kyttä (1976), and Kyttä and Hurme (1982) review the properties attributed to vowel nasalization in the literature. It is clear that descriptions vary widely, but the authors group the observations into four main categories: a weakening of total intensity (owing to the weakening of individual formant intensities and a broadening of their bandwidth); a relatively weak first oral formant; the introduction of nasal formants at 250, 1000 and 2000 Hz with anti-formants in between, and associated shifts in oral formant frequencies; and some minor irregularities at high frequencies.

The effects of nasalization can be illustrated with examples from Fant's (1970) study using an electrical vocal tract analogue. The low back vowel [aa] produced by his model had formants at 630, 1070, 2400 and 3550 Hz. With a small degree of coupling (0.16 cm sq.), the only changes visible in the spectrum were a splitting of F3 into two peaks and a valley (by the addition of a pole-zero pair) and a raising in frequency of F4. When the coupling was increased

to 0.65 cm sq., F1 fell to 600 Hz, F2 narrowed in bandwidth, F3 was as for the oral vowel, but of lower amplitude, and F4 was of higher amplitude. With an increase in coupling to 2.6 cm sq., F1 remained at 600 Hz and widened in bandwidth, F2 decreased in amplitude and became much wider in bandwidth, and F3 was lowered in frequency; an extra formant appeared *above* F1, between 800 and 900 Hz, and a zero separated F3 and F4. The vowel [ii], on the other hand, showed a *rise* in F1 frequency (along with an increase in its bandwidth) from 220 to 250 Hz, and an additional pole-zero pair above 1000 Hz (a peak at 1100 Hz and a valley at 1800 Hz), while the higher formants remained unchanged in frequency.

3.5. The acoustic properties of the vocal tract cavities

It was seen in sections 3.2 and 3.4 that the production of nasality can involve all the cavities of the vocal tract. There is, in fact, no speech sound in which the effects of the nasal tract are present in isolation: even in velar and uvular nasal stops the excitation passes through the pharynx before reaching the nasal cavities, while in the nasal release of oral stops, when sound energy is released into the nasal tract at the velopharyngeal port itself, the pharynx is coupled through the open port. It is necessary, then, to consider the acoustic properties of all these cavities when examining the nasal spectrum.

3.5.1. The acoustic properties of the nasal tract

The acoustic properties of the nasal tract are still a matter for debate. Particular problems are the separability of the response of the nasal cavities

themselves from that of the nasopharynx, and the role of the sinus cavities.

According to Delattre (1969a: 95, quoted by Cagliari 1978: 192), only the nasopharynx

> has walls firm enough for efficient resonance; the other nasal cavities, those which terminate in the nostrils, have fibrous walls and could only have a damping effect.

Other writers (e.g. Fant 1970, Kaplan 1971) accept that the nasal cavities proper are indeed capable of resonance, but views differ as to the nature of their response: Kaplan sees the nasal cavities as forming a multiple resonator, while Fant holds that the upper, middle and lower passages are too closely coupled to function as independent resonators. All writers tend to agree that the nasal cavities' response is highly damped, owing to the high ratio of the circumference of the cavities to their cross-sectional area at any one point except in the nasopharynx (Bjuggren and Fant 1964) – a consequence of their complex internal structure referred to in section 3.3 – and the presence of the nostril hairs at the outlet (Fant 1970: 141).

It appears that the first natural resonant frequency of the nasal *cavities* proper lies near 1000 Hz. This figure comes from a modelling study by Bjuggren and Fant (1964), and is based on measurements of the nasal cavities taken from a plaster cast of a male speaker, treating the cavity as a Helmholtz resonator bounded by the nostrils anteriorly and the choanae posteriorly.* There is nothing in the literature about the resonant frequency of the nasopharynx, partly because its response depends on the position of the velum. According to

* There is some doubt here, however, since the posterior coordinate of the nasal cavities in their Figure 1-B-2 – coordinate 9 – appears to take in part of the nasopharynx.

Bjuggren and Fant (1964) it is capable of functioning as an independent resonator in slight degrees of nasalisation – that is, with a small amount of velopharyngeal opening. Otherwise, its independence is lost: with complete closure of the velopharyngeal port, it functions as part of the nasal tract (Bjuggren and Fant 1964); while with full velopharyngeal opening – as in nasality – it functions as part of the pharynx.

The first resonant frequency of the whole nasal tract, including the nasopharynx, appears on the basis of modelling studies to be between 400 and 600 Hz (House and Stevens 1956, House 1957, Hecker 1962, Bjuggren and Fant 1964, Lindqvist and Sundberg 1976), with an inverse relationship to the *length* of the nasal tract model. Fant (1985) gives the rather lower figure of 297 Hz for a model which includes the effects of two pairs of nasal sinuses. Estimates of the frequencies (and number) of the higher resonances vary considerably. House and Stevens (1956) and House (1957) found a single higher peak at 2500 Hz (with maximal velic coupling). Hecker (1962) found a peak varying between 1300 and 2000 Hz, depending on the degree of coupling, and an *anti-resonance* between 600 and 1300 Hz. Lindqvist and Sundberg (1976), using direct excitation of the closed nasal tract by a swept-frequency sound source inserted into the nasopharynx, found peaks at around 600, 1000, 1250 and (two peaks) just above 2000 Hz for one male speaker, and at around 700, 900, 1100, 1600 and 2200 for another. They also found zeros, but these were related to the location of the sound source in the nasopharynx: when the sound source was moved from 2 cm from the velum to around 5 cm from the velum the zero frequency

fell from over 3000 Hz to just under 2000 Hz. This frequency presumably reflects the resonance of the volume posterior to the sound source; in the latter case, with the source 5 cm from the end, it is probable that the source no longer lay in the nasopharynx, but inside the nasal cavity proper (given the measurements presented by Bjuggren and Fant 1964: 5), and this frequency might therefore represent the natural frequency of the nasopharynx.

Fant (1970) demonstrates the effect of velopharyngeal coupling on the relationship between the nasal cavities and the nasopharynx: a small coupling area gave peaks at 500 and 2000 Hz (my estimates from his Figure 2.4-3c), showing both the fundamental resonance of the whole nasal tract and the contribution of the nasopharynx, while a larger coupling area, which presumably allowed the nasopharynx to function with the pharynx instead of with the nasal cavities, gave resonances at 0, 1000, 2800 and 4200 Hz.

The effects of asymmetry in the nasal cavities, such that one cavity has a greater cross-sectional area or length than the other (as observed by Bjuggren and Fant 1964), have been largely discounted: most studies of the nasal tract use either simple unbranched models, or two-tube models with perfect symmetry. Its main effect seems to be to cause additional complexity in the nasal spectrum, since the nasal output becomes a mixture of the responses of two slightly different resonators. According to Fant,

> asymmetry will cause an additional diffusion of spectral energy owing to the occurrence of formants from the left and the right pathways, and to the particular mixing in the nasal radiation

(1970: 141). Lindqvist and Sundberg (1976) estimate that any asymmetry

would introduce an extra pole-zero pair between each resonance. Fujimura and Lindqvist (1964a) attribute the complexity seen at low frequencies in their sweep-tone study of nasal stops to this phenomenon, but also suggest that it could be due to the paranasal sinuses.

3.5.2. The paranasal sinuses

The role of the sinuses in nasal cavity resonance is unclear. Many writers have asserted that they make no contribution (e.g. Van Riper and Irwin 1958: 245; Zemlin 1968: 252). Reasons given for this view are that the cavities involved tend to be fluid-filled, rather small and with small or easily blocked openings into the nasal cavity. Greene apparently allows for some contribution by the maxillary sinuses, since these are large and "open into the nose by fairly large orifices" (Kaplan 1971: 225, quoting Greene). According to Kyttä (1970) blocking of the maxillary sinuses with radio-opaque material (to allow its position to be checked) produced no significant changes in spectrograms, though these probably lack the necessary sensitivity.

Their resonance characteristics have not been measured directly, since the cavities are practically inaccessible. Some estimates have been made, however, on the basis of their size, assuming that they behave as Helmholtz resonators (e.g. Lindqvist and Sundberg 1976, Maeda 1982, Fant 1985). Lindqvist and Sundberg (1976) estimate their natural frequency at between 200 and 800 Hz for the maxillary sinuses and between 500 and 2000 Hz for the (rather variable) frontal sinuses. They do not give the source of the anatomical measurements on which they base these calculations. Each sinus pair would have only

one resonance (seen in the nasal tract transfer function as one pole-zero pair) in the frequency range studied, so long as its members were identical in size; any asymmetry would introduce *three* pole-zero pairs corresponding to each sinus pair. Fant (1985) suggests a resonance frequency of 399 Hz for the maxillary sinus pair and 1399 Hz for the frontal sinuses as suitable frequencies for nasal tract modelling. Lindqvist and Sundberg point out that exact estimates of the resonances of the sinuses are unnecessary for modelling, given the wide variation seen among speakers in both nasal cavity and sinus dimensions.

Their resonance function would be a complex one, since they would form closed side-branches to the nasal tract, thereby contributing both anti-resonances and resonances to the transfer function of the nasal cavities. How these features appeared in the output spectrum would depend on whether the nasal tract formed the main branch (as in nasal stops) or the side-branch (as in nasal vowels).

Two studies have suggested that their involvement is essential for accurate modelling of the nasal tract frequency response: in one case (Lindqvist and Sundberg 1976) to account for the complexity observed at low frequencies in swept-frequency studies (Fujimura and Lindqvist 1964a, Lindqvist and Sundberg 1976), and in another (Maeda 1982) to allow a vocal tract model to generate a low enough first formant in the production of nasalised vowels. Castelli and Badin (1988) also allow the possibility that the sinuses may be responsible for the low frequency of the nasal resonance seen in their white-noise excitation of the vocal tract during the production of velar nasal stops.

In the study by Lindqvist and Sundberg, the addition of two shunting cavities intended to correspond to the maxillary and frontal sinus pairs gave a good match to the observed sweep-tone response curves for the nasal tracts of their subjects, with peaks at 500, 900, 1250 and 2000 Hz (my estimates from their Figure 5b), but neither the dimensions nor the specific acoustic characteristics of these shunting cavities are given. Asymmetry in the sinuses was not explored.

Maeda's study (1982) compared the output of a simple three-tube model of the vocal tract system, having pharyngeal, oral and nasal branches, with a model containing an additional side-branch simulating the contribution of the maxillary sinus pair. Without the sinus cavity, the lowest resonance of the isolated nasal tract in the model was 670 Hz; the addition of the sinus – having a volume of 20.8 cm cubed and a coupling section (neck) 0.5 cm long and 0.1 cm sq. in area – introduced a single pole-zero pair to the nasal tract transfer function, splitting the original 670 Hz peak into peaks at 446 and 817 Hz, with a zero between at approximately 500 Hz, close to the 550 Hz natural frequency of this cavity. The higher frequency nasal resonances were not affected. The use of this nasal tract model in generating the nasalized vowel [aa] did seem to improve the detail of the spectrum (his Figure 4), introducing a second pole-zero pair below that introduced by the nasal cavity itself (his Figure 6), and thereby lowering the frequency of the observed "nasal formant" below that of the shifted oral formant.

It is reasonable to assume that the sinuses do make a contribution to the response of the nasal tract, then, but the significance of this response is not clear. Modelling studies need to be viewed with caution, since the acoustic properties of the sinuses depend heavily on the dimensions assumed in the model, and these can easily be chosen to give results which bear out the investigator's hypothesis. The addition of the sinuses to the nasal tract can be presumed to lower its fundamental resonance in much the same way as coupling of the oral cavity to the nasal-pharyngeal tract lowers the resonance of the vocal tract, but it would be wrong to attribute the main low-frequency resonance of nasal stops, for example, to the sinuses themselves, since no matter how low their own resonances may be, they are closed cavities with no communication to the outside air and their output would be severely attenuated compared with that of the nasal tract itself. In addition, it is difficult to see how their frequency could be lower than that of the nasal-pharyngeal tract, with its much greater length and volume.

3.5.3. The nostrils

The cavities formed by the external framework of the nose at the nostrils are small and highly damped, and have not been attributed with any resonance function in the production of nasality except in their role as outlets from the nasal cavities proper. Alteration of the nostril area has important acoustic effects in the transfer function of the nasal tract, however (see 3.5.6).

3.5.4. The oral cavity

In the production of nasalized vowels, the oral cavity functions much as it does for oral vowels, as part of the main acoustic tube with the pharynx. In the production of nasal stops, it functions as a closed side-branch to the main acoustic passage, which is here formed by the pharynx and the nasal tract (Fant 1970).

The acoustic properties of the oral cavity in the production of nasalized vowels are much the same as those involved in the production of *oral* vowels, except for the effects of the lowering of the velum on the area function at the back of the mouth (Maeda 1982) (which for some writers, in any case, constitutes part of the pharynx, the oral cavity ending at the faucal pillars: e.g. Romanes 1986: 137). These effects are considered negligible by some writers (e.g. House and Stevens 1956, who make no provision for them in their vocal tract analogue), but others state that the changes are important for correct vowel quality in synthetic speech (e.g. Maeda 1982).

The properties of the oral cavity in nasal stops are more interesting. It contributes the main anti-resonances observed in the spectrum of stops (Fant 1970, Fujimura 1962, 1963), the frequencies of these anti-resonances reflecting the natural resonant frequencies of the oral cavity and depending on the location of the stop constriction and the behaviour of the tongue body under the influence of adjacent vowels.

According to calculations made by Fujimura (1963) the first natural resonant frequency of the oral cavity is usually not lower than 700 Hz, being

higher for closures further back in the mouth. The second lies at around 3000 Hz or higher for [m], and much higher for [n]. These figures relate to the cavity *behind* the oral constriction, since this is what is relevant for nasal stop production. The possible contribution of the cavity to the *front* of the constriction will be considered below.

Fujimura calculates that a simple uniform acoustic tube with a length of 8 cm, representing the oral cavity during the production of [m], would cause an anti-resonance at around 1000 Hz, higher in the case of coarticulation with a high front vowel (in which the tongue body would narrow the mouth tube anteriorly) and lower in the case of coarticulation with a back vowel (in which a large cavity is produced with a rather narrow posterior opening). An oral cavity tube of length 5 cm would give an anti-resonance at 1700 Hz – a representative figure for the alveolar nasal [n]. A tube of length 3 cm would represent the oral cavity "shunt" for the velar nasal [ng], giving an anti-resonance above 3000 Hz.

Fujimura's calculations are partly based on an interpretation of the results of a spectral matching experiment, using a system of poles and zeros to fit the observed spectrum of real speech tokens. They agree with the observations of Fant (1970) on the output of an electrical analogue based on anatomical measurements. Fant regards the oral cavity during the bilabial nasal stop as a Helmholtz resonator, closed at the lip end, having a fairly large volume and a relatively narrow neck formed by the passages to the pharynx on each side of the uvula (which touches the tongue at the midline). In his calculations, this

configuration (with a neutral tongue position) has a natural resonant frequency of around 800 Hz. An increase in coupling to the pharynx, by incomplete lowering of the velum or a lowered tongue position, would cause a rise in this frequency to 1000 Hz.

The response of the mouth cavity also depends on the position of the tongue body, which varies according to the phonetic identity of neighbouring vowels. Coarticulation with [ii], for instance, as in Fant's measurements of the Russian palatalized nasal /m_j/ (1970: 147) causes the natural resonant frequency to rise to around 1800 Hz, giving rise to a zero at that frequency in the output spectrum.

Characteristics of the front oral cavity

The characteristics of the cavity in *front* of the oral constriction (in the case of closures posterior to the alveolar ridge) have not been considered. This cavity is apparently assumed to play no role in the production of nasal stops, since there is usually no coupling with the cavity *behind* the constriction, which is what determines the frequency of the major anti-resonance; if there were such coupling, with incomplete closure between the tongue and the palate, for example, we might expect this anti-resonance to fall in frequency with the enlargement of the side-cavity, and also perhaps for oral cavity formants to appear in the output spectrum. It has been suggested that the velar and uvular nasal stops may have front oral cavity resonance despite this lack of coupling, and that this can be demonstrated

> by producing either of them with a strong whisper, and changing the position of the lips from a spread posture to a rounded one and back again. The pitch of the resonances of the front of the mouth can be quite clearly heard, falling markedly with increasing lip-rounding and rising again with progressive lip-spreading

(Laver 1980: 84). This resonance would presumably arise from transmission of sound energy through tissues such as the tongue and soft palate. There is some acoustic evidence for such transmission in Fujimura and Lindqvist (1971: 551), who noted the occurrence in sweep-tone spectra for the *oral* velar stop [g], "particularly when there is appreciable lip-rounding, ... [of] a peak that apparently shows the resonance of the mouth cavity in front of the tongue". Such a peak can be seen in the spectrum of [g] before the vowels /o/ (at 800 Hz) and /u/ (at 600 Hz), while no peak is seen when this stop precedes the unrounded back vowel /ɯ / (their Figure 11). However, this effect has not been reported in any acoustic studies of nasal stops.

Hattori and co-workers (1958) attributed a second anti-formant seen between 2000 and 3000 Hz in the spectrum of [m] coarticulated with the front vowels [ii] and [e] to the front oral cavity.

3.5.5. The Pharynx

The pharynx is relatively simple in its structure by comparison with the nasal tract, but it also is highly variable and its acoustic properties are therefore difficult to define absolutely. If it is treated simply as an acoustic tube of constant cross-section, its resonances will be determined solely by its length: an adult male pharynx of length 10 cm (from the glottis to the inferior surface of the raised velum, Fant 1980), for example, would have resonances at odd-numbered multiples of the fundamental resonance of 875 Hz according to

Equation A.33-5 in Fant (1970: 292):

$$F_n = (2n-1)c/4l, \tag{3.2}$$

where F_n is the frequency in Hz of resonance n and c is the speed of sound in normal speaking conditions (approximately 35300 cm/s). However, its length and cross-section are highly dependent on the position of the tongue and velum. A very low position of the velum, with a large coupling to the nasopharynx, means that the nasopharynx becomes part of the pharyngeal tube, increasing its length and thereby lowering both its fundamental resonance and the interval between resonances.

Fant (1970) suggests that in the production of nasal stops, the pharynx can also be regarded as a Helmholtz resonator, with the nasal tract as the resonator neck; in this case a fundamental resonance nearer to 300 Hz is obtained (Fant 1970: 142).

There are also numerous small cavities within the pharynx whose resonance function is not often made explicit. The paired recesses on either side of the larynx, the *sinus piriformis*, for example, contribute a zero just above 5000 Hz, according to Fant (1970: 102), but their effect is generally ignored or treated as one of several constant factors contributing to high frequency attenuation.

3.5.6. The nasal spectrum and the contribution of cavities

Various writers have attributed features of the nasal spectrum to the contributions of individual cavities. Kyttä (1970), for example, attributes the first formant of the nasal stops (between 200 and 250 Hz) specifically to the

"epihypopharyngeal cavity" (suggesting the entire pharynx, including the nasopharynx but excluding the nasal cavities proper), the second formant (between 1300 and 1600 Hz) to the "epipharyngeal cavity" (that is, the nasopharynx); and one formant each to the three nasal *meati* – inferior, medial and superior – at 1950-2300 Hz, 2300-2700 Hz and 2700-3300 Hz respectively. Fant, however, emphasizes the combined contribution of the nasal cavity and other parts of the vocal tract. For example, he attributes the first formant of the velar nasal to the fundamental Helmholtz resonance of the pharynx "tuned by the nasal system as a resonator neck" (1970: 142), and observes that the second formant of the velar nasal [ng] in his model (at 1000 Hz) is close to the calculated impedance minimum of the nasal tract and is therefore presumably due to that minimum. The third formant, at 2200 Hz, is "mainly related to a half-wavelength resonance of the pharynx", and the fourth, at 2900 Hz, to the second impedance minimum of the nasal cavity system.

This more cautious approach is well justified, since resonance is the result of the *entire* vocal tract configuration, not of any single cavity within it. As Fant observes,

> all parts of the vocal tract contribute to the determination of all formants but with varying degree depending on the actual configuration

(1970: 21). A similar point is also made by House (1957: 202):

> If the vocal system is viewed as a number of simple independent or loosely-coupled resonators, it would be expected that these resonances would in some way be characteristic of nasal vowels and consonants. On the other hand, in a closely-coupled system of resonators, the normal modes of vibration of the component parts are not the normal modes of the system as a whole. When nasal consonants are produced, the characteristics of these sounds cannot be deduced from a knowledge of the transfer characteristics of the nose itself, since the

> various resonators comprise a closely-coupled system and interactions take place among the system components. As pointed out by Dunn [Dunn, H.K. (1950), J. Acoust. Soc. Amer. 22, 740-753 – *EJR*] and others, assigning vowel formants, for example, to specific cavities is not a valid procedure ...

This interdependence of cavities means that the observed features of nasal stops – even those which are apparently attributable to the nasal cavities – are influenced not only by the structure and shape of the cavity mainly responsible for their production, but also by the state of the cavities to which they are coupled, and by the nature of the coupling itself.

The effects of coupling changes on the response of the nasal tract are demonstrated by Fant (1970: Fig.2.4-3c): a reduction in the coupling area to the nasal tract model causes a lowering of its resonance frequencies by several hundred Hz. Fant also reported that the shunting effect of a side-cavity varies according to its coupling to the main acoustic path: at large coupling areas, the pole-zero pair which it contributes to the spectrum are well separated in frequency, but as the coupling area is reduced the zero moves closer to the pole until, with complete closure, they neutralize each other. This applies to both the nasal tract during the production of nasal vowels, and the oral cavity in the production of nasal stops: Fant notes, for example, that increased coupling between the pharynx and the oral cavity in [m], by a lower tongue position, can shift the oral anti-resonance from around 800 Hz up to 1000 Hz, possibly neutralising the "nasal cavity" resonance. In the case of the nasal tract response, the situation is complicated by the changing relationships between the nasal cavities and the nasopharynx with changes in coupling (Bjuggren and Fant 1964).

In nasal stops, then, we can expect the formant frequencies contributed by the nasal tract to vary with changes in velopharyngeal coupling. While such changes in coupling have not been observed directly, changes in velic *height* have been seen (cf. 3.2.8), though whether they reflect coupling changes is unclear. A perceptual study by Abramson and co-workers (1981) suggests that these differences may be perceptible, since a greater degree of velopharyngeal opening was required for a synthetic alveolar stop to be judged as nasal when it occured before an open vowel.

The effects of a coupling change on a cavity's response also depend on the coupling of the cavity in question to *other* cavities, however. Thus Fant (1970: 152) observes that the effects of varying the *nostril* area in his model depended on the degree of *velopharyngeal* coupling, and that with a large degree of coupling the lowering in frequency of the nasal tract zeros was smaller. The same will hold for the oral cavity contribution to nasal vowels, where there is coupling between the oral cavity and the pharynx at the fauces, and between the oral cavity and the outside air at the lips.

Changes in coupling also have the effect of altering the balance between the oral and nasal contributions in the output spectrum, in the case of nasal vowels. House (1957), for example, observes a general rise in the acoustic impedance of the nasal cavities over all frequencies as the coupling area (in the model) is reduced. This means that the acoustic features of the oral cavity output (including nasal cavity anti-resonance effects) will predominate in the observed signal, and nasal cavity resonance (and oral anti-resonance) effects

may be obscured. Conversely, Fant (1970) notes that nasalized close vowels (in which the coupling area between the oral cavity and the pharynx is reduced) show greater energy in the nasal output than in the output from the mouth.

Changes in coupling apart, alterations to the other vocal tract cavities themselves will also have an effect on the nasal spectrum because they affect the vocal tract configuration as a whole. Thus while an increase in the volume of the oral cavity will lower the frequency of its anti-resonance in nasal stops (Fujimura 1962, 1963), as was observed in the preceding section, it will also increase the total volume of the vocal tract as a whole, lowering its fundamental resonance (Fant 1970: 145). This will affect not only the first formant of the nasal stops, which is dependent on all cavities, but also all other resonances to some extent.

The problem of predicting the acoustic effect of particular gestures is made more difficult by the fact that a given gesture may have various physical consequences. Changes in velopharyngeal coupling, for example, achieved by movements of the velum, have an effect on the response of the nasal tract as indicated above. However, they almost certainly alter the area function at the back of the oral cavity as well, affecting its own response *and* its coupling to the pharynx. As mentioned above, the effects of these particular changes are generally regarded as insignificant (e.g. House (1957) and House and Stevens (1956)), though Maeda (1982) suggests that the inclusion of such complementary changes are important for good quality modelling.

3.6. Acoustic consequences of cavity changes

It is clear from the preceding section that the frequency characteristics of the vocal tract cavities are not constant. This section examines how changes in these characteristics caused by changes in cavity dimensions may be reflected in the output spectrum of speech.

3.6.1. Nasal cavity changes

It will be remembered that the only changes which the nasal cavities are capable of undergoing are a narrowing or widening of the passages on either side of the median septum by the erectile tissue, and blockage of all or any of the meati (again on either side) by mucus (3.3). The acoustic effect of any narrowing is presumably a raising of any resonance frequencies which depend on the overall volume of the nasal cavities, but resonances which depend on the nasal cavities functioning as the neck of a Helmholtz resonator (the fundamental resonance of the velar nasal stop, for example: Fant 1970) will show a fall in frequency. Fant (1970) predicts that, conversely, in a nasal tract which is "especially wide and free from constrictions", the first nasal zero in the spectrum of nasalized vowels may be found as high as 1800 Hz, and even above the second vowel formant.

Very few studies describe the effects of such changes, however. Sambur (1975: 180) shows the effects of a mild head cold (presumably accompanied by nasal congestion) on the Linear Prediction spectrum of an alveolar nasal stop: his Figure 4 shows a rise in all formant frequencies by several hundred Hz. This is only in a single token, however, and it is not clear how much of this

change is attributable to nasal cavity changes alone. Lindqvist and Sundberg's (1976) experience with the effects of irritation of the nasal linings during sweep-tone analysis of the nasal tract response suggests that the nasal secretions may reduce the *complexity* of the nasal cavity response: when adrenaline was applied to the nasal mucosa on the side through which the sound source was inserted (to *reduce* the secretions), the spectrum changed drastically, having many more peaks below 1000 Hz.

The only other attempts to monitor their acoustic consequences have involved artificially stopping up the nasal passages in some way (Hattori et al. 1958, Kyttä 1970). Both studies used spectrographic analysis, so their results are not as detailed as we might wish. Hattori et al. (1958) observed the effects of local constrictions in the nasal cavities on the frequency of the anti-resonance of nasalized vowels. A bilateral stoppage at the nostrils lowered the anti-resonance location, but as the blockage was moved further inside the nasal cavities, the frequency rose, and was highest when the blockage was located at the "inner points" (undefined). At this point, the anti-resonance frequency was very dependent on the vowel articulation; the authors attribute this to the variation in the configuration of the nasopharynx (on which the anti-resonance frequency apparently depended) with the level of the tongue. Kyttä (1970) reports a similar experiment, in which the dental nasal stop [n̪] was pronounced with various blockages to the nasal cavities. Complete closure on one side (using cotton-wool soaked in brine) at an unspecified point gave "a general weakening of energy in the higher frequencies" (96). With this unilateral

blockage still in place, the first formant of the nasal remained at 250 Hz so long as the lowest nasal passage on the other side was free; when the upper and middle passages were obstructed, the "resonance of the upper nose passages" (between 2000 and 3300 Hz, according to Kyttä) was weakened or disappeared altogether, while the 250 Hz formant was reinforced. With a blockage of the lowest passage while the upper passages were free, the 250 Hz formant was weakened (but no information is given about changes to its frequency), while the higher frequency energy remained. Simple closure of the nostrils caused the first formant to rise to 350 Hz; the further back the bilateral blockage was placed, the higher the formant rose, until

> when the midway to the septum has been reached, i.e. approximately 4 cm from the nostrils, the formant rises up to 500-1000 cps, which is the same level as the [oral] anti-resonance has. The nasalization stops, and the nasal consonant in question changes into some other sound

(1970: 98) – most commonly, something resembling the corresponding oral stop.

Nasal congestion can also be expected to block the entrances to the sinuses and to lead to accumulation of mucus within them. This may not prevent them from resonating (much as blockage of the nasal cavities does not prevent their resonance), but it can be expected to shift their resonances (by a reduction in their volume or a change in the area of the opening forming the resonator neck) and to damp their contribution to the nasal spectrum even more severely.

Kyttä (1970) investigated the effects of artificially introduced sinus blockages using spectrographic analysis, but found no significant changes. Castelli and Badin (1988) speculate that a sinus blockage could be responsible for

differences seen in one of their subjects between the frequency response of the right and left nostrils, when the vocal tract was excited by a white-noise source: the spectrum for the right nostril shows fewer clear peaks than that for the left nostril, with the peaks at 600-700 Hz and 1500 Hz (my estimates from their Figure 6, p.420) being more severely attenuated.

3.6.2. Effects of changes in the oral and pharyngeal cavities

Apart from changes in the nasal tract itself, the major source of change in the vocal tract response is the variation in the oral cavity and pharynx resulting from jaw and tongue body movements for vowel articulation. The pharynx is also susceptible to larynx movements, which correlate with tongue movements to some extent in vowel articulations, and also with fundamental frequency changes.

Such movements naturally occur constantly in connected speech, even during the articulation of nasals, and Glenn and Kleiner's assertion that "the articulators do not move during the period of oral closure" (1968: 368) is not true, as Fujimura (1962) has demonstrated.

In nasal stops, the oral cavity is chiefly responsible for the frequency of the major anti-resonance, and changes to its shape and volume can be expected to affect this feature. Fujimura (1962, 1963) predicts that the oral cavity anti-resonance of [m], for example, will rise in the case of coarticulation with a high front vowel (in which the tongue body would narrow the mouth tube anteriorly) and fall in the case of coarticulation with a back vowel (in which a large cavity is produced with a rather narrow posterior opening). This effect is

demonstrated by Fant (1970) with data from a vocal tract analogue. Palatalization of "neutral" [m] and [n] (equivalent to coarticulation with the vowel [ii]) produces a rise in anti-resonance frequencies from 800 and 3500 Hz to 1800 and 5600 Hz for the bilabial nasal, and from 1800 and 5600 Hz to 2200 and 6400 for the alveolar nasal, with a small effect on the formant frequencies. Hattori et al. (1958) noted the occurrence of an extra anti-formant in the region from 2000 to 3000 Hz in the case of bilabial nasals coarticulated with the front vowels [ii] and [e].

Changes in the oral configuration take place continuously during normal speech, so some movement of the oral cavity anti-resonance can be expected. Fujimura (1962) illustrates the effects of tongue movement on intervocalic [m] and [n] during the transition from a neutral vowel to the vowel following the nasal stop. During the [m] of [h@'miim] the antiformant rose in frequency from 800 to 1200 Hz, carrying with it the second and third formants (as a formant-antiformant "cluster"), while the first and fourth formants remained almost unchanged. In the [n] of [h@'naan], the downward and backward movement of the tongue was accompanied by a fall in the frequency of the antiformant (and associated formants) from 1700 to 1500 Hz.

The damping of the oral cavity response, and therefore the bandwidth of the anti-resonance seen in the nasal output, also varies with the mouth configuration. Fujimura points out that the alveolar nasal [n] shows a higher bandwidth anti-formant because the wedge-shaped anterior termination to the cavity, where the tongue blade meets the alveolar ridge, causes greater acoustic

losses than the abrupt closure at the lips for [m].

The response of the pharynx has been shown to be quite sensitive to variations in its length caused by larynx movement, at least in the production of vowels (Sundberg and Nordstrom 1976, Ashby 1983). No studies have been made, however, of the effects on nasal stops. Larynx height variations of as much as 2 cm were recorded in running speech by Ashby (1983), with corresponding changes of up to 8% in either direction in the second formant frequency of [ii] (closely related to the length of the back cavity, including the pharynx). These changes were correlated with fundamental frequency movements. The position of the larynx also varies with vowel articulations, though, by virtue of the attachment of both the tongue body and the larynx to the hyoid bone. A correlation between vowel height and larynx height has been observed in some studies, while Laver mentions

> momentary positional fluctuations [in larynx height] caused by the movements of the muscles directly involved in, or passively affected by, the production of segmental articulation

(1980: 26). Voluntary movements of the larynx, such as those produced by Lindqvist and Sundberg's subjects, are also possible, and can give changes of several hundred Hz in formant frequencies in the long term spectrum (Laver 1980: 29).

3.6.3. Vowel coarticulation effects on nasal spectra: acoustic data

Coarticulatory effects of vowel context on nasal spectra have received relatively little attention in acoustic studies of real speech, apart from the limited examples referred to in the last section. Kurowski and Blumstein, for example,

state that "there are no reported tables of values for English nasal resonances in labial and alveolar consonants across various vowel contexts" (1984: 385), and with the exception of the data they themselves give (see below), and two abstracts by Oyer and co-workers (1986a, 1986b) dealing with the bilabial nasal only, little or nothing has been published since. Velar nasals in particular have been neglected: only one study (Saito and Itakura 1984) deals with them (see Chapter Six), but the vowel context is fixed. One reason for this neglect is that most interest lies in *anticipatory* coarticulation effects, which in English can be seen only in bilabial and alveolar nasals, since only these occur syllable initially. Another reason is that the main effect in alveolar and bilabial nasals appears to be on the location of the principal anti-resonance, which in velar stops apparently lies at too high a frequency to be of interest to some researchers.

What acoustic evidence there is suggests that the spectrum of bilabial nasals shows the greatest variation with vowel context, presumably as a result of the freedom of the entire tongue to coarticulate with neighbouring vowels (Nolan 1983, Repp 1986). Su, Li and Fu (1974), for example, showed this using multivariate analysis of speech spectra for bilabial and alveolar nasal stops in four speakers. Principal components analysis was used to project the spectral samples of stops in front vowel and back vowel contexts onto just two dimensions: for all four speakers, samples from [m] preceding front vowels were clearly separated from samples from [m] preceding back vowels, but this was not the case for [n]. The Euclidean distance between the centroids of the two

distributions was around three times as great for the bilabial stop as for the alveolar stop.

Kurowski and Blumstein (1984) give formant frequencies for 5 vowel contexts derived from LPC analysis of word-initial bilabial and alveolar nasal stops in isolated CV syllables for a single male speaker. The frequencies (in Hz) of the first 4 formant peaks visible in the steady-state portion of the nasal stop (from their Table 1) are given in Table 3.1. No clear trends are visible, but it can be seen that the number of peaks detected varies considerably with vowel context. No information is given on *anti-resonance* frequencies, but since the authors used Linear Prediction, this would be of limited use.

Oyer and co-workers (1986a, 1986b) also used Linear Prediction to derive formant frequencies for the bilabial nasal stop in a variety of contexts. The first formant frequency was found to be influenced by vowel context, especially

Nasal stop	Vowel	N1	N2	N3	N4
m	i	247	2079	3199	-
	e	252	1236	2118	3335
	a	251	852	2608	3393
	o	246	807	2591	3369
	u	240	922	2033	3186
n	i	237	1766	2532	3342
	e	251	800	1723	2567
	a	256	839	2524	3373
	o	251	797	2560	-
	u	239	1382	2543	3421

Table 3.1 Formant frequencies of LPC spectra for a single male speaker (from Kurowski and Blumstein 1984)

in initial position in monosyllables, but no acoustic data are given, and again there is no information on anti-resonances. The authors conclude by observing that

> a comprehensive study of the sounds subject to the least coarticulatory effects may prove to be important in speaker identification

(1986b: s61).

Several authors state that vowel coarticulation effects on the velar nasal spectrum are minimal (e.g. Fujimura 1962: 1869), and this is supported to some extent by Castelli and Badin (1988) who quote Feng (1986) as finding that velar nasal stops show very little dependence on a coarticulated vowel; they also mention that in their own data (from white-noise excitation of the vocal tract during velar nasals coarticulated with various vowels by two male speakers), formant frequencies vary by less than 5%. Again, however, no information is provided on anti-resonance frequencies.

3.6.4. Summary

Changes in the characteristics of the nasal spectrum have indeed been observed with variation in both the nasal and oral-pharyngeal tracts, but published data allowing this variation to be quantified are rather limited. Nasal cavity changes are particularly poorly understood, but even the effects of vowel coarticulation, which can be controlled much more easily, have received relatively little attention. What studies there are concentrate exclusively on resonance data, and information on anti-resonances is not presented, despite the fact that the major influence on the nasal spectrum to be expected from vowel

effects is on the anti-resonance contributed by the oral cavity. Data on the velar nasal stop are practically non-existent. There is considerable scope, then, for further work on the acoustics of nasal stops.

3.7. Summary: the variability of the nasal spectrum

We have seen in this chapter that the nasal tract is involved in the production of nasality, as part of the primary channel in the production of nasal stops and as a side-branch in the production of most nasalized segments including vowels and fricatives, and that it *may* be involved in the production of auditory nasality as a long-term quality or setting.

There is evidence too for the considerable inter-speaker variability in nasal cavity size and structure, commented on by several researchers. This variability is indeed seen in the acoustic properties of the nasal tract.

Nasality is more than simply resonance of the nasal cavities, however: they only ever operate in conjunction with other cavities of the vocal tract, notably the pharynx cavity and the oral cavity, and there is no single acoustic feature exclusively dependent upon their configuration. Their contribution to the speech waveform varies with the configuration of these other cavities, and with changes in coupling, changes which occur constantly throughout speech, even in nasal stops. In addition, though the configuration of the nasal cavities themselves is not consciously variable, it is subject to fairly rapid change – not only from day to day but from hour to hour. The effects of such changes on the nasal spectrum have never been quantified.

Despite these reservations, though, nasality offers the only feature in speech which depends on a part of the vocal tract whose structure cannot be changed at will. If it is to be used for speaker verification, however, care must be taken to maximize the contribution made by the nasal cavities to the speech wave. It is suggested that nasal stops provide the most suitable environment, having the greatest degree of velopharyngeal opening of any speech segment (Cagliari 1978), and, simultaneously, the highest possible oral cavity impedance.

Even with nasal stops, however, the effects of changes in the oral cavity and pharynx introduce unwanted variation into the spectrum, and the selection of an articulation which minimizes these changes would be beneficial. This leads to the choice of articulations towards the back of the oral cavity, that is at the *velum*or *uvula*: here the oral cavity branch is at its smallest, so that its anti-resonance effects are restricted to higher frequencies, and the tongue is highly constrained in its movements compared with labial and lingual-coronal articulations. For speaker verification in British English, then, the choice comes down to the *velar* nasal stop /ng/, and it is this segment which is chosen for investigation in the remainder of this thesis.

A secondary reason for the choice of this segment is that it is the velar place of articulation which has received least attention in both automatic speaker recognition work (Chapter Two, 2.8.2) and acoustic studies of nasality itself (3.6.3). Several studies have observed that velar nasal stops do indeed show a greater resistance to the effects of coarticulation than bilabial and

alveolar nasal stops, but there is little published in the way of acoustic data. In particular, the changes in anti-resonance frequencies found in the velar nasal have not been dealt with, nor have anti-resonance features been considered for use in speaker verification.

The following chapters concentrate, then, on the velar nasal stop, looking at the nature and extent of its acoustic variability in terms of both poles and zeroes, and assessing the potential for automatic speaker verification of some suitable representations of the velar nasal spectrum. Chapter Four and Chapter Five consider the choice of an appropriate analysis method for finding pole and zero frequencies; in Chapter Six this method is used to obtain data on pole and zero frequencies in velar nasal stops for a large group of speakers, examining vowel context effects and variation over time; while in Chapter Seven the pole-zero spectral data are used in a series of verification trials to demonstrate that the velar nasal has considerable potential for speaker verification.

CHAPTER FOUR

POLE-ZERO DECOMPOSITION OF SPEECH SPECTRA

CHAPTER FOUR

POLE-ZERO DECOMPOSITION OF SPEECH SPECTRA

4.1. Introduction

In this Chapter, the suitability of methods for characterising nasality in terms of both poles (resonances) and zeroes (anti-resonances) is considered. The need for such a method emerged in the discussions of nasality presented in Chapter Three, in which the velar nasal stop was chosen for investigation. These discussions suggested that though the effects of *oral cavity* anti-resonance on the spectral structure of the velar nasal should be minimal (this being one of the reasons for its selection), some anti-resonance effects are nevertheless to be expected. These anti-resonances may result from

(1) asymmetry in the nasal passages (3.5.1);

(2) the contribution of the paranasal sinuses (3.5.2);

(3) incomplete velar closure (that is, between the tongue dorsum and the lower surface of the velum), leading to possible coupling with the oral cavity (3.5.4);

(4) the possible resonance and anti-resonance effects of the front oral cavity, even in the case of complete velar closure (3.5.4);

(5) the coupling of the sub-glottal cavity to the vocal tract during the open phase of the glottal cycle. (Closed phase analysis (Steiglitz and Dickinson 1977) is one way to remove this influence, but it may not be effective on speakers with breathy or whispery voice for whom there is always some sub-glottal coupling.)

This chapter therefore considers a variety of methods for estimating the spectral structure of speech, and looks at their ability to handle the presence of zeros (anti-resonances) as well as poles. Two general approaches – *all-pole modelling* and *pole-zero modelling* – are outlined, and the superiority of pole-zero modelling for use particularly with nasal segments is demonstrated. The discussion then concentrates on the method for pole-zero decomposition in the cepstral domain proposed by Yegnanarayana (1981). An experiment with synthetic speech is used to show the suitability of this method for the purposes of this thesis.

4.2. Modelling speech production

The vocal tract can conveniently be viewed as a time-varying linear filter, with either the periodic glottal waveform or random noise as input and the speech waveform itself as output (Atal 1985: 81). The difference between the input waveform and the speech waveform is in this view solely the result of the filtering action of the vocal tract.

This view of speech production – known as the *source-filter* model – has proved extremely useful in speech research. It is summarised succinctly by Fant (1970: 15):

> **the speech wave is the response of the vocal tract filter function to one or more sound sources**

and is represented schematically in Figure 4.1 (a).

The sound source or *excitation* in the model is provided by the periodic vibration of the vocal folds for *voiced* speech, or by the creation of random (aperiodic) noise for *voiceless* speech. The filtering action of the vocal tract modifies the frequency content of the source spectrum by reinforcing some frequencies and absorbing others. The resulting speech output has a frequency spectrum which shows peaks (formants) at frequencies corresponding to the natural resonant frequencies of the vocal tract in its current state, and dips at frequencies corresponding to the anti-resonances of the vocal tract – that is, the frequencies at which the vocal tract absorbs energy.

The aim of much speech analysis is to estimate the frequency response of the vocal tract filter to allow the extraction of features such as formant frequencies and bandwidths, and to characterize the excitation by its fundamental frequency and amplitude. This can be done using a variety of frequency domain techniques, including filter-bank analysis and the Discrete Fourier Transform. Alternatively, a mathematical model of the filter characteristics may be derived. Such a model summarizes the transfer function of the vocal tract in an economical way, and can be used to generate synthetic speech.

The use of mathematical modelling of the vocal tract transfer function is perhaps the most widespread method of speech analysis. In essence, the filtering action of the vocal tract during a short interval, when it can be assumed to

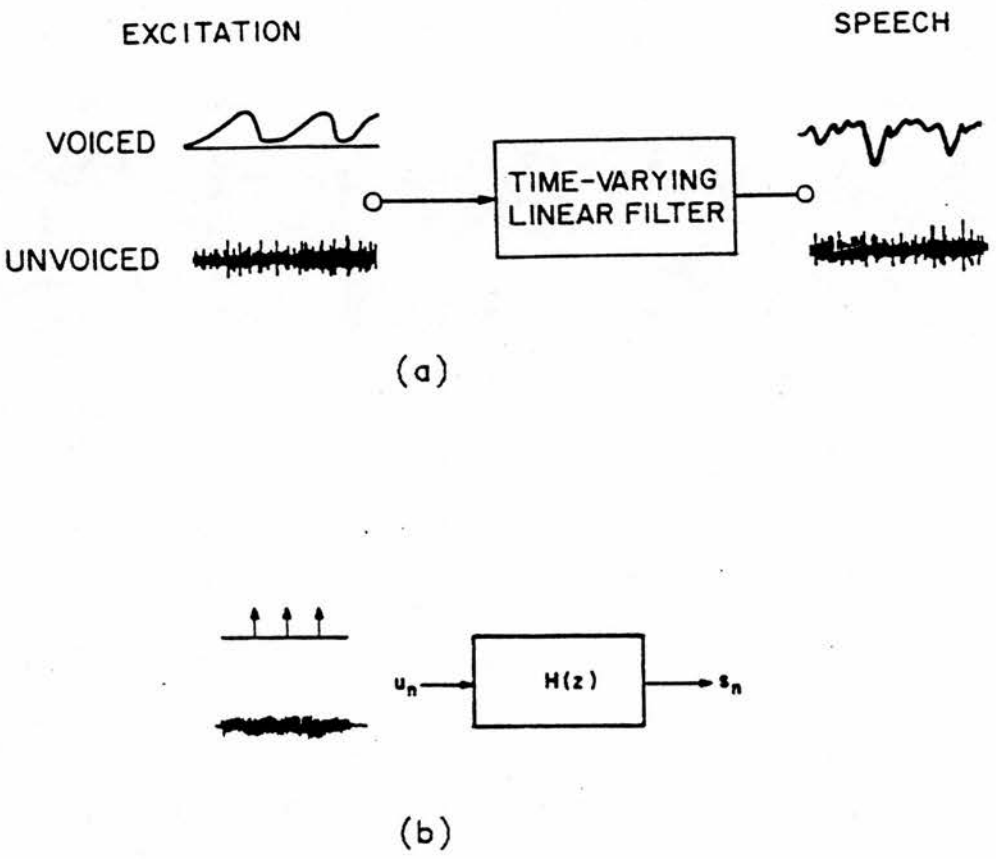

Figure 4.1 Source-filter models of speech production (from Atal 1985) (a) basic model (b) simplified, discrete model

be stationary, is imitated by an expression built up from polynomials – that is, a *digital filter*. Appropriate coefficients are chosen to build up the filter, so that its output, when excited by either a periodic or random noise source, approximates the observed speech waveform over this interval. The resulting filter therefore represents the filtering action of the vocal tract, and can be used to derive smoothed frequency spectra and formant characteristics for speech analysis purposes, or to synthesize a version of the original speech by exciting

it with a suitable source function.

Underlying most mathematical modelling is a simplified version of the source-filter model of speech production, shown diagrammatically in Figure 4.1 (b). The source is assumed to be an impulse train (spikes of energy, each of negligible duration), with an amplitude and fundamental frequency equal to those of the observed (or synthesized) speech waveform, or a section of white noise (that is, completely random vibration). Each of these sources has a completely flat frequency spectrum, with components of equal amplitude at all frequencies, so the spectral shape of the output is solely the result of the action of the filter. Without constraining the source (input) in some way, it would not be possible to derive a filter to match an observed speech waveform. However, this particular constraint means that the filter represents not only the frequency response of the vocal tract (the goal of the analysis), but also the spectral characteristics of the excitation (which does not have a truly flat spectrum), the effects of radiation of the speech wave from the lips and any effects introduced by the environment or recording equipment used to obtain the observed speech waveform. This considerably simplifies the derivation of the model.

The dominant method of modelling is Linear Predictive Coding (LPC), in which the filter consists of a set of coefficients (predictor coefficients) which "predict" the observed speech signal by a weighted linear combination (a sum) of past outputs and present and past inputs (i.e. values of the excitation function). The equation representing such a filter is as follows:

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G\sum_{l=0}^{q} b_l u_{n-l} \tag{4.1}$$

where s_n represents the observed sequence of speech samples, and u_n represents the input or excitation sequence. This filter can also be represented in the frequency domain using z transform notation (Oppenheim and Schafer 1975: 45) as follows:

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1+\sum_{l=1}^{q} b_l z^{-l}}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{4.2}$$

where S(z) is the z transform of s_n and U(z) is the z transform of the input sequence u_n (Makhoul 1975: 562, Equation 1). H(z) as defined here is known as the *general pole-zero model.* The roots of the numerator polynomial S(z) (that is, the coefficient values which will give zero output) are the *zeros* of the model, while the roots of the denominator polynomial U(z) (that is, the coefficient values which give infinite output) are the *poles* of the model. The poles and zeros may be real or in complex conjugate pairs.

Two special cases of the model are obtained when either the numerator or denominator coefficients are set to zero. Setting all the numerator coefficients b_l (where $1 \le l \le q$) to zero gives an *all-pole* model: its output is a linear combination of the present input value, weighted by the gain value G, plus the weighted values of past outputs. This model is known in the statistical literature as the autoregressive (AR) model (Makhoul 1975: 563):

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{4.3}$$

In the frequency domain, this can be expressed in z transform notation as:

$$H(z) = \frac{G}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{4.4}$$

– an all-pole transfer function. Setting all the denominator coefficients a_k (where $1 \leq k \leq p$) to zero gives an *all-zero* model, whose output is a linear combination of the present input value with weighted past input values. This model is known in the statistical literature as the *moving average* (MA) model. The combined pole-zero model is therefore known as the *autoregressive moving average* or ARMA model (see for example Moir 1988).

The *all-pole* model is of special interest in speech analysis because it can be used to represent the frequency response of a single acoustic tube (Atal 1985: 81), such as is formed approximately by the unconstricted oral-pharyngeal vocal tract, or that of a set of resonators connected in sequence. The all-pole model is also relatively easy to derive, and therefore forms the basis of practically all time-domain speech modelling procedures.

4.2.1. All-pole modelling: Linear Predictive Coding

The technique of Linear Prediction, while in theory capable of using the general pole-zero model, is almost always based on the all-pole model.

Linear prediction (LP, or Linear Predictive Coding, LPC) is a time-domain coding technique applicable to all discrete time-based signals (such as EEG signals or seismic survey data, for example (Makhoul 1975)), but of great importance for speech modelling. Each frame of digitized speech – usually between 10 and 25 milliseconds long, during which time the vocal tract can, except for transient sounds, be assumed to be held constant – is reduced to a small set of

parameters: the gain (or amplitude) of the source, its fundamental frequency if periodic, and a set of predictor coefficients representing the frequency response of an all-pole vocal tract filter. This form of representation is ideal for transmission and storage, and is of equal use for analysis and synthesis. Its chief advantage in analysis is that it separates the parameters of the source (the gain and fundamental frequency) from those of the filter (the predictor coefficients or their equivalent frequency response) without the need to go into the frequency domain as in Fourier analysis. This facilitates the extraction of frequency domain information such as formant values, which are obscured by the presence of the harmonic "ripple" on Fourier spectra. It also provides many alternative representations which are easily related either to frequency-domain concepts such as the power spectrum and cepstrum or to other mathematical models of the vocal tract's behaviour such as *area functions* and *reflection coefficients* of acoustic-tube models of speech production (e.g. Fant 1970). In addition, it is relatively easy to implement, assuming as it does both a simple model of speech production and an all-pole mathematical model.

Selection of the coefficients of the model could be made on several criteria, but the usual requirement is that the chosen coefficients should minimize the difference between the observed signal and the output of the model. This is a reasonable criterion for successful modelling, but it also simplifies the derivation of the coefficients, since the resulting minimisation problem reduces to a set of simultaneous linear equations (Makhoul 1975: 563), which are easily solved using matrix algebra to give the coefficients a_k. This is one of the

advantages of the all-pole model.

Solving the Linear Prediction problem in this way also has the consequence that the frequency spectrum of the error signal (or residual) is maximally flat (Witten 1982: 129). This is so irrespective of whether the original speech was voiced or voiceless. If the speech was voiced, minimizing the total (or mean) squared error gives an error signal which approximates an impulse train (the excitation assumed by the model for voiced speech); if the speech was voiceless, the error signal approaches white noise (the excitation assumed for voiceless speech). Both of these signals have a spectrum which is flat – that is, has components of equal amplitude at all frequencies.* Thus, simply by solving for the set of coefficients which minimizes the total error, Linear Prediction models both the source function (given by the residual) and the transfer function (given by the predictor coefficients).

The choice of the *number* of coefficients in the model is important. The greater the number (that is, the higher the model *order*), the more faithful will be the model's representation of the observed speech waveform; as a consequence, the frequency spectrum of the response of the resulting filter will be much more like that of the original speech. If the model order becomes too high, however, the frequency spectrum of the model will begin to show the "ripple" which results from the periodic excitation (Atal 1985: 91), and the major advantage of Linear Prediction – its separation of source and filter information – will have been lost. If the model order is too low, the model will be inaccu-

*The spectrum of the residual will never be completely flat, since a small number of predictor

rate. Markel and Gray (1976) have shown that, as a minimum, the "memory" of the model (that is, the total delay in sample points, or equivalently, the number of coefficients) should be equal to twice the time required for sound waves to travel from the glottis to the lips; that is, $2L/c$, where L is the length of the vocal tract and c is the speed of sound. For a vocal tract of length $L = 17$cm, and a representative value of $c = 34$ cm/millisecond for the speed of sound in the vocal tract, the memory must be at least 1 millisecond: that is, 10 sample points at a sampling rate of 10 kHz, requiring a model order of 10 coefficients.

It is also necessary to add coefficients to the model to take account of the fact that both the glottal wave spectrum and the lip radiation characteristics are included in the vocal tract modelling (since the source function is assumed to have a flat spectrum). Finally, a few coefficients are added owing to the fact that the digitized speech waveform is not exactly an all-pole waveform, even when no *vocal tract* zeros are involved (Markel and Gray 1976: 154). A reasonable compromise, suggested by Markel (1971b), is that the model order should equal the sampling rate in kHz (the minimum) plus 4 or 5 coefficients to accommodate the above-mentioned factors. This solution allows for the modelling of one formant in every 1 kHz of the spectrum up to the Nyquist frequency (half the sampling frequency), since two poles are needed to model a single formant; this is adequate for most phonetic analyses.

---

coefficients cannot be expected to model the speech waveform perfectly.

Linear prediction analysis produces a set of predictor coefficients as a model of the vocal tract filter characteristics. This is a time-domain representation, but in phonetic analysis, as opposed to speech synthesis or transmission, we are usually interested in frequency-domain information such as formant locations and bandwidths or spectral slope. One way of obtaining such information is to solve for the *roots* of the model, the poles of the denominator polynomial (e.g. McCandless 1974). These are the frequencies at which the model's output tends to infinity. Root-solving is, unfortunately, rather complex and requires a high degree of mathematical precision (Yegnanarayana 1978). An alternative method is to calculate the frequency response of the model directly from the coefficients, by Fourier analysis.* This gives a smoothed log magnitude (or power) spectrum, from which measurements of formant frequency and bandwidth can be made by relatively simple peak-picking routines.

4.2.2. Pole-zero modelling

4.2.2.1. The need for pole-zero modelling

All-pole modelling has been so successful because it is simple, is well-understood, can be carried out easily in the time domain and gives good results in the analysis of most speech sounds, particularly where re-synthesis is the goal. The all-pole model it uses is not applicable to all speech sounds, though: it can only accurately model the behaviour of an *unbranched* resonator (the vocal tract during vowel production, for example). However, the production of

*This is possible because the predictor coefficients define the model's *impulse response* – that is, its output when excited by a single non-zero sample value.

nasals involves the introduction of a side branch to the main acoustic tube (Fant 1970); the production of *laterals* may involve a splitting of the air-passage on either side of a central constriction (Catford 1977: 132); while the production of voiceless *fricatives* at places of articulation above the glottis involves a sound source within the vocal tract, with a resonant cavity behind as well as in front of the constriction (Flanagan 1972). All these conditions may introduce zeros into the vocal tract transfer function, corresponding to anti-resonances in the spectrum. In addition, in voiceless fricatives, during the open phase of each glottal cycle, or continuously where incomplete glottal closure persists as in breathy or whispery voice, sound energy is absorbed by the cavities below the glottis causing a widening of the bandwidths of the vocal tract resonances (Flanagan 1972: 65) and perhaps the introduction of a significant sub-glottal anti-resonance (Nolan 1983: 155; Fujimura and Lindqvist 1971). Zeros can also be introduced by the use of a low-pass anti-aliasing filter before digitization of the recorded speech signal during analysis.

Zeros in the transfer function of speech (whether originating from vocal tract anti-resonances or from sub-glottal interaction) have two effects on the spectrum: the introduction of a local dip and an alteration to the spectral balance (Atal 1985: 82; Atal and Schroeder 1978: 1311). For example, the digital pre-emphasis filter frequently used in speech analysis

$$x(n) = x(n) - \mu x(n-1) \tag{4.5}$$

consisting of a single zero:

$$H(z) = 1 - \mu z^{-1} \tag{4.6}$$

where μ is the pre-emphasis factor (normally between 0.9 and 1.0), has the effect of modifying the whole spectrum so that higher frequencies are preferentially emphasized over lower frequencies. It is possible to approximate some of the effects of a zero on the spectrum using an all-pole model, by including extra pole coefficients (Atal and Hanauer 1971; Makhoul 1975: 577). This is often done in Linear Prediction analysis with satisfactory results. However, only the effects on the spectral balance can be modelled satisfactorily in this way: local dips in the spectrum are very difficult to model using extra poles without introducing ripples at other frequencies in the spectrum (Atal and Schroeder 1978). In addition, if there are several zeros in the input spectrum, the number of poles required becomes very large (Makhoul 1975: 577).

It has been suggested that the inclusion of zeros in Linear Prediction for synthetic speech gives an improvement in perceived quality, at least for nasals (Mermelstein 1972; Atal and Schroeder 1978), although Atal and Schroeder also found that some benefit could also be achieved through the use of a higher-order all-pole model.

Where the details of the spectrum of a particular sound segment are of interest, however, the inclusion of zeros in the model is highly desirable, particularly when the location of the spectral zeros themselves is important, as in the study of nasality.

4.2.2.2. The problems of pole-zero modelling

The inclusion of zero coefficients in the model is not a simple matter, however. While the choice of coefficients in the all-pole model can be determined

by solving a set of *linear* equations to find the solution which gives the minimum error, attempts to solve for pole and *zero* coefficients simultaneously lead to a set of *non-linear* equations with no simple solution to give the single minimum error (Makhoul 1975: 576-577). Methods do exist to overcome this problem, but they are much more complex than the algorithms available for the all-pole solution and less well understood. Many are also untested on real speech waveforms.

4.2.2.3. Solutions to the problem

Solutions to the problem of non-linearity in minimizing the error to find pole and zero coefficients are of two types (Makhoul 1975: 577): *iterative* and *non-iterative*. Iterative methods make repeated estimates of the coefficients of the model, evaluating the error and adjusting the parameters accordingly. Such methods are time-consuming and, unless initial estimates of parameter values are fairly accurate, may never converge to give a reasonable solution (Markel and Gray 1976: 273). Non-iterative methods generally require a good estimate of the *number* of poles and zeros in the input signal if they are to give accurate results. Even where this is not possible, non-iterative methods can be used to give reasonable initial estimates of the coefficients as a starting point for an iterative solution (Makhoul 1975). Most methods reported in the literature (and reviewed below) are non-iterative.

4.3. Pole-zero modelling methods

4.3.1. Iterative methods

Very few iterative methods have actually been applied in speech analysis, because of the time they take to converge (and the possibility of non-convergence).

Steiglitz and McBride (1965) describe a technique which they call *iterative prefiltering*, applicable to any linear system, not just speech. A block diagram of the technique is shown in Figure 4.2. An initial estimate of the model parameters (the numerator coefficients $N(z)$ and the denominator coefficients $D(z)$) is obtained using a method suggested by Kalman (1958) for the minimiza-

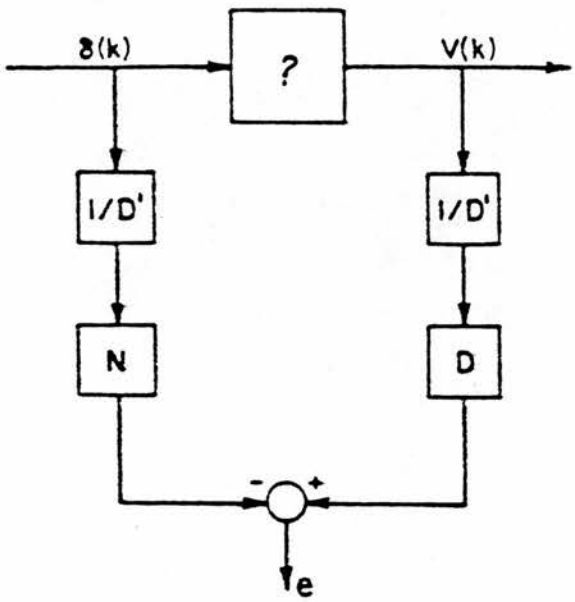

Figure 4.2 The iterative prefiltering method of pole-zero modelling (Steiglitz and McBride 1965)

tion of the error between the input and output sequences of the system. The initial estimate of the denominator $D_1(z)$ is then used to *pre-filter* the input and output sequences, and new estimates of $N(z)$ and $D(z) - N_2(z)$ and $D_2(z) -$ are obtained from the pre-filtered sequences. Each new estimate of $D(z)$ is used to filter the original input and output sequences to give new pre-filtered sequences for input to the next iteration. The process is continued until the error reaches a minimum. The authors report that, using a synthetic signal, between 10 and 20 iterations were needed to derive a third or fourth order model. No tests were carried out on speech. Presumably, the number of iterations required would be much larger for higher order models such as are required for speech modelling.

Steiglitz (1977) compares an implementation of iterative prefiltering with one non-iterative method for solving the non-linear minimization problem (Shanks 1967). In this comparison, the initial estimate of $D(z)$ used for prefiltering the input was obtained by Linear Prediction, instead of using Kalman's (1958) method. Linear Prediction was also used to derive the estimate of $D(z)$ for use in Shanks' non-iterative method. A single frame taken from the bilabial nasal [m] was analysed using the two methods; the resulting spectra are shown in Figure 4.3. A 14th-order LPC analysis gave an initial estimate of the denominator coefficients. The spectrum of this all-pole model, shown in Figure 4.3 (b), has poorly-modelled spectral dips and too-sharp spectral peaks (that is, the bandwidths of the formants are too narrow) compared with the (DFT-derived) smoothed log magnitude spectrum in Figure 4.3 (a).

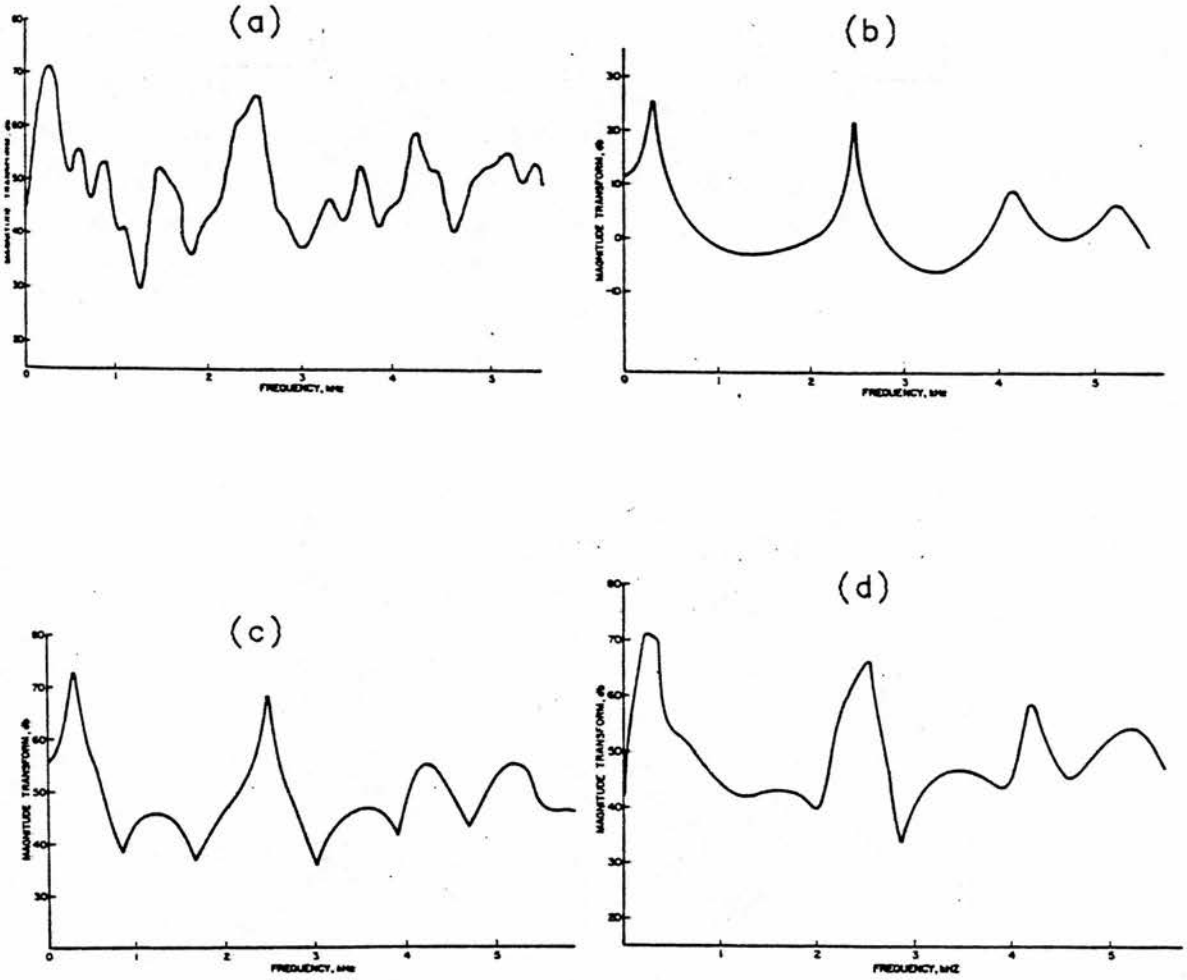

Figure 4.3 Magnitude spectra of the bilabial nasal [m] (from Steiglitz 1977): (a) cepstrally smoothed DFT spectrum (b) smoothed Linear Prediction spectrum (c) pole-zero model (Shanks 1967) (d) pole-zero model (iterative pre-filtering)

Shanks' method was then used to obtain an estimate of the numerator coefficients (19th order), given the 14th-order denominator polynomial estimated by LPC. The spectrum of the resulting pole-zero model (Figure 4.3

(c)) shows better resolution of the spectral dips, but again has narrow estimates for the bandwidth of the formant peaks, since the pole parameters are still those given by the original all-pole LPC analysis. Application of the iterative prefiltering method, with new (19th order) numerator and (14th order) denominator polynomials being estimated at each iteration, gave the spectrum shown in Figure 4.3 (d) (after ten iterations): while the spectral dips are rather inconsistently modelled, the peaks are very accurately represented.

A more recent attempt at an iterative method is described by Moir (1988). This constructs a pole-zero Linear Prediction model using a technique for recursive parameter estimation known as extended recursive least squares (Ljung and Södeström 1982). No evaluation of the method is given, however, other than that it gives "similar results to the all-pole method but with a reduced order of predictor" (1988: 1355).

4.3.2. Non-iterative methods

Iterative methods overcome the non-linearity of the solution to the minimization problem by attempting repeated solutions with revised pole and zero parameters, while monitoring the error produced. *Non-iterative* methods generally change the problem to a related linear one by fixing some of the parameters – usually the denominator coefficients – to allow the calculation of the remaining parameters using linear methods. Most therefore are "two-pass" algorithms: the denominator coefficients are estimated using all-pole modelling (as for LPC), and the numerator coefficients estimated separately.

An early attempt at a non-iterative solution was suggested by Shanks (1967). His solution has two stages, each involving the solution of a linear problem (see Figure 4.4). The first step is to estimate the denominator coefficients D of the model by applying a form of linear prediction (using the Covariance method) to an estimate of the *impulse response* corresponding to the observed signal, rather than to the signal itself. This is done by solving the minimization problem:

$$\min_{D} \sum_{k=M+1}^{K} \left[Dv(k) \right]^2 \tag{4.7}$$

(Steiglitz 1977: 230, Eqn.6); Shanks' method then finds the set of numerator coefficients N which minimizes the error between the model's output and the impulse response given the estimate of the denominator:

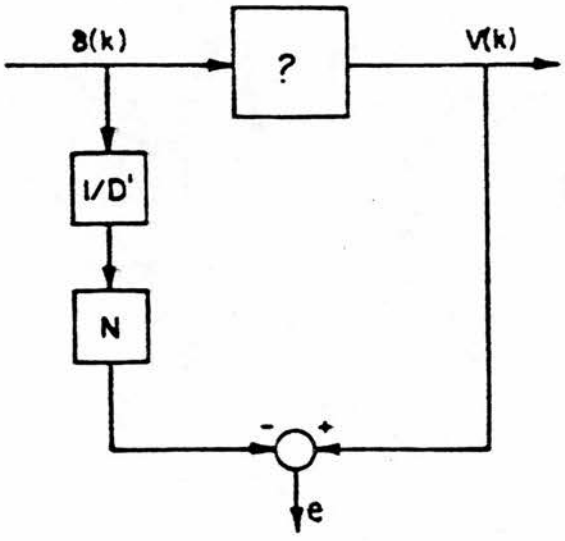

Figure 4.4 Shanks (1967) non-iterative method of pole-zero modelling

$$\min_{N} \sum_{k=0}^{K} \left[v(k) - \frac{N\delta(k)}{D'} \right]^2 \tag{4.8}$$

(Steiglitz 1977: 230, Eqn.8). At least $2M$ samples are required to do the analysis where M is the model order. This method of estimating the denominator and numerator coefficients separately is common to several of the techniques presented in this section.

The method proposed by Kalman (1958), and used by Steiglitz and McBride (1965), also starts with the impulse response $V(k)$, but attempts to estimate N and D simultaneously by solving the linear problem:

$$\min_{N,D} \sum_{k=0}^{K} \left[Dv(k) - N\delta(k) \right]^2 \tag{4.9}$$

(equation 9 from Steiglitz 1977: 230).

Like all non-iterative methods, these methods both require as a starting point an estimate of the *impulse response* of the system, or equivalently a single time-synchronized pitch period (Markel and Gray 1976: 272). The remaining algorithms presented here provide ways of estimating this impulse response by filtering the input spectrum using Linear Prediction or homomorphic (cepstral) analysis.

Makhoul (1975), reporting on work published in 1974, describes a technique for modelling zeros by first transforming them into poles, to allow them to be described by an all-pole Linear Prediction filter in the usual way. The method, which he terms *inverse linear prediction*, relies on the removal of the estimated poles from the input signal before the calculation of the zero coefficients.

The pole coefficients are derived first using an autocorrelation approach. The resulting all-pole filter is inverted, and the signal passed through it to remove the effects of the spectral poles. The all-zero spectrum is then low-pass filtered (by autocorrelation, cepstral smoothing or all-pole modelling), inverted (to make it all-pole), and subjected to all-pole modelling by linear prediction. The resulting pole coefficients are "good estimates" of the numerator coefficients of the pole-zero model.

Atal and Schroeder (1978), reporting work first published in 1975, describe another method based on Linear Prediction with spectral inversion to convert zeros into poles.

They derive an estimate of the model's impulse response by a two-stage Linear Prediction, and then estimate the pole and zero coefficients simultaneously by solving a set of linear equations. Figure 4.5 shows a block diagram of their approach.

The impulse response (which includes not just the filtering action of the vocal tract but the effects of lip-radiation, the slope of the spectrum of the glottal source and — something rarely acknowledged elsewhere — the spectral contribution of the recording procedure) is derived by a high-order LPC analysis (25th order at a sampling frequency of 10 kHz) of just two pitch periods using a modified covariance method. No additional windowing (beyond the rectangular window implied by the selection of the input data) is applied. The resulting prediction error signal or residual is modelled by a second 25-pole LP analysis, which effectively inverts its all-zero spectrum. The predictor coefficients

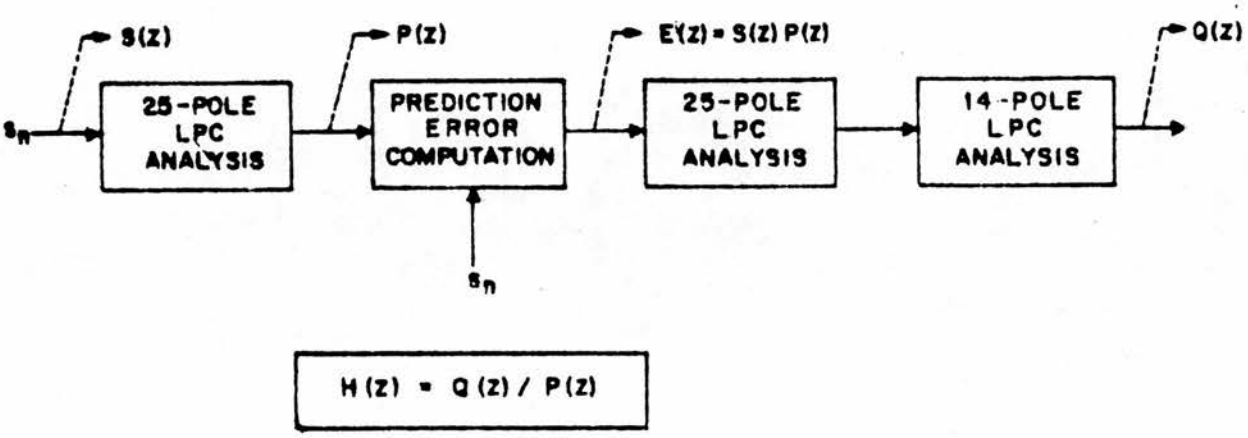

Figure 4.5 Linear Prediction pole-zero analysis (Atal and Schroeder 1978)

derived at this stage represent a spectrum which is the approximate inverse of the spectral envelope of the error signal. A third LP analysis, this time 14th order, inverts *this* spectrum, producing a set of coefficients Q which represent the envelope of the spectrum of the predictor error. The impulse response itself is obtained by dividing this polynomial $Q(z)$ by the polynomial produced by the very first LP analysis ($P(z)$). The pole and zero coefficients of the model are then obtained from the impulse response by solving a set of linear equations.

Atal and Schroeder point out that their method of obtaining the pole parameters is similar to that of Shanks (1967), but that their method of calculating the zero coefficients is different.

4.3.3. Homomorphic prediction

The term *homomorphic prediction* has been applied to any of a set of techniques in which an estimate of the vocal tract impulse response is obtained by cepstral analysis (homomorphic filtering) before Linear Prediction is applied to generate the parameters of a pole-zero model (Kopec et al. 1977). Two very similar methods are described in the literature.

Oppenheim and Tribolet (1973) describe a technique which they call "cepstral prediction", in which the numerator and denominator of the model are derived by two all-pole Linear Prediction analyses. The basic principle is that the poles of $nC(n)$, the low-quefrency weighted complex cepstrum, comprise the poles *and zeros* of the observed signal. Once the roots of the model have been found, via Linear Prediction analysis, it is necessary to decide which are poles and which zeros. To avoid this, the poles are first estimated by a Linear Prediction analysis on the observed signal. The all-pole filter which results is used to filter this signal, and the complex cepstrum is derived; Linear Prediction on the weighted low-quefrency portion then gives the zero coefficients (Markel and Gray 1976: 275; Makhoul 1975: 578). According to Markel and Gray (1976), although the technique works well for synthetic speech, experiments with real speech have not been so successful.

Kopec and co-workers (1977) describe an algorithm in which the vocal tract impulse response is estimated by cepstral smoothing of the DFT spectrum. The signal is pre-emphasized, the FFT spectrum calculated and the cepstrum derived. The low-quefrency components are selected to give an estimate of the

vocal tract impulse response (equivalent to the smoothed spectrum in the frequency domain). Linear Prediction is then applied to this signal to estimate pole coefficients. The minimum-phase impulse response signal is then *inverted* and passed through the all-pole filter; this removes its "zeros", that is, the poles of the uninverted signal, leaving a signal corresponding to the effects of the spectral zeros.

This signal is then modelled by Linear Prediction to derive a set of pole coefficients. Because they are derived from the inverted all-zero impulse response, they correspond to the zero coefficients of the original smoothed spectrum.

In commenting on this method, Markel and Gray (1976: 273) point out that it has difficulty resolving nearly coincident pole-zero pairs.

Yegnanarayana (1981) presents a technique which involves decomposition of short-time spectra in the cepstral domain, followed by a form of linear predictive modelling to derive pole and zero coefficients.

It exploits a relationship between the cepstrum and the *negative derivative of phase* spectrum (NDPS). This spectrum has the important property that its positive peaks can be attributed largely to the influence of complex conjugate poles, while its negative peaks (dips) can be attributed to the influence of spectral zeros. The NDPS can be calculated from the cepstral coefficients by the relationship:

$$\Theta_V(\omega) = \sum_{k=1}^{\infty} c(k) \sin k\omega \qquad (4.10)$$

(Equation 10 from Yegnanarayana 1981).

By selecting only the low-quefrency cepstral coefficients for this operation, the NDPS of an estimate of the smoothed vocal tract impulse response is obtained (that is, the filtering action only, with the excitation removed). The positive and negative real values of the NDPS are separated, and separate positive and negative cepstral responses are derived by the inverse of the above calculation. Thus the cepstral response for the poles of the vocal tract filter model is effectively deconvolved from the cepstral response for the zeros. Each cepstral response is then modelled by an all-pole filter using a form of Linear Prediction, to give the numerator (zero) coefficients and the denominator (pole) coefficients of the complete pole-zero model. The frequency response of each part of the model can be calculated by a Fourier transform on the relevant coefficients, and the combined model frequency response is given by the arithmetic sum of the separate pole and zero frequency responses.

One difference between this technique and others presented above is that the estimate of the zero coefficients is independent of the estimate of the pole coefficients, being made "in parallel", rather than after removing the (estimated) pole activity from the signal. It is perhaps more likely to be able to resolve overlapping poles and zeros in the spectrum than are other approaches.

4.3.4. Summary

The method proposed by Yegnanarayana (1981), being based on the cepstrum rather than on the Linear Prediction all-pole model, looks the most promising for the analysis of nasals because the use of the all-pole model is

inherently inaccurate. The two-pass algorithms (Makhoul 1975, Atal and Schroeder 1978, Oppenheim and Tribolet 1973, Kopec et al. 1977), even those based on cepstral analysis rather than Linear Prediction, all suffer from the disadvantage that the estimate of the zero activity in the signal is dependent on an accurate prior estimate of the pole activity, since the signal from which the zero activity is estimated is derived by filtering with the inverse of the all-pole spectral model. One consequence of this is that it is necessary to estimate the *number* of poles required in advance. If too high a model order is used, some of the waveform behaviour caused by spectral zeros will be modelled by the poles and the estimate of *both* will be inaccurate.

Finally, the ability to separate closely spaced pole and zero pairs is an important feature for a study of nasality. It is doubtful whether the two-pass algorithms outlined above possess this ability (Markel and Gray 1976: 273).

The remainder of this chapter therefore concentrates on a detailed description of the Yegnanarayana (1981) algorithm, and assesses its suitability for the analysis of nasality by using synthetic speech tokens.

4.4. The cepstral decomposition method of pole-zero modelling

The method proposed by Yegnanarayana (1981) was summarised in the last section. It is considered in greater detail here, and in the following section its response to a synthetic controlled signal will be investigated.

4.4.1. Discussion of the method

4.4.1.1. General description

Yegnanarayana's technique amounts to a decomposition of the cepstrum – the Inverse Fourier transform of the log power spectrum – into separate pole and zero components. This decomposition is achieved by exploiting a property of the phase spectrum. The *phase* spectrum is derived from the complex Fourier spectrum; it expresses the relationship between the sine and cosine components which make up the *magnitude* spectrum, and without it the frequency analysis is strictly speaking incomplete. Where the magnitude of each component is calculated from the complex spectrum as the square root of the sum of the squared real and imaginary parts

$$|X(f)| = \sqrt{Re(f)^2 + Im(f)^2} \tag{4.11}$$

(where $|X(f)|$ is the magnitude of the spectrum at frequency bin f), its phase angle Θ is given by the arctangent of the *ratio* between the imaginary and real parts:

$$\Theta(f) = \tan^{-1}(Im(f)/Re(f)) \tag{4.12}$$

Phase is often disregarded in frequency analysis, partly because a sound wave reconstructed from just the magnitude spectrum (that is, with the components of the phase spectrum all effectively set to zero) sounds the same to the human ear (Fant 1970, p.235).

It has been shown that the phase spectrum alone can be used to reconstruct a signal under certain conditions (Hayes et al. 1980, Quatieri and Oppenheim 1981; both cited by Yegnanarayana 1984): that is, it contains

similar information about the frequency spectrum to that given by the magnitude, but in a different form.

In an earlier paper (1978), Yegnanarayana shows, in particular, that the *derivative* of the phase spectrum provides an alternative to the magnitude (or power) spectrum for the estimation of formant frequencies and bandwidths in speech analysis. The *derivative* of the phase spectrum (DPS) can be obtained simply by differentiating the phase spectrum: that is, calculating the difference between adjacent samples. According to Yegnanarayana (1978), extraction of formant information from the DPS is a promising alternative to the usual method using the magnitude or power spectrum, owing to a crucial property of the DPS. This is that the overall phase spectrum is a *summation* of the contributions of individual resonators, rather than a *multiplication* as in the magnitude spectrum. Individual resonators (or alternatively complex pole pairs in the transfer function) do not affect each other's phase response, therefore. This means that closely spaced poles (or formant peaks) will be easier to resolve from the DPS than from the magnitude spectrum.

Yegnanarayana (1978) uses this property to propose a method for estimating formant frequencies from Linear Prediction spectra: a 12th order Linear Prediction analysis on a speech segment gives a set of predictor coefficients representing the smoothed frequency spectrum; a 512 point DFT on these coefficients generates the complex frequency spectrum; the phase component of this spectrum is obtained, and then differentiated, smoothed to remove the effects of phase *wrapping,* and inverted. The resulting *negative derivative of*

phase spectrum (NDPS) has peaks which can be used to give formant frequencies (from their location) and bandwidths (from their height).

Figure 4.6 illustrates this for a segment of the neutral vowel [@]. The log magnitude LP spectrum (c) and the negative derivative of phase spectrum (b) were calculated from a 512 point DFT on the predictor coefficients produced by a 24th order LPC analysis. The close correspondence between the two

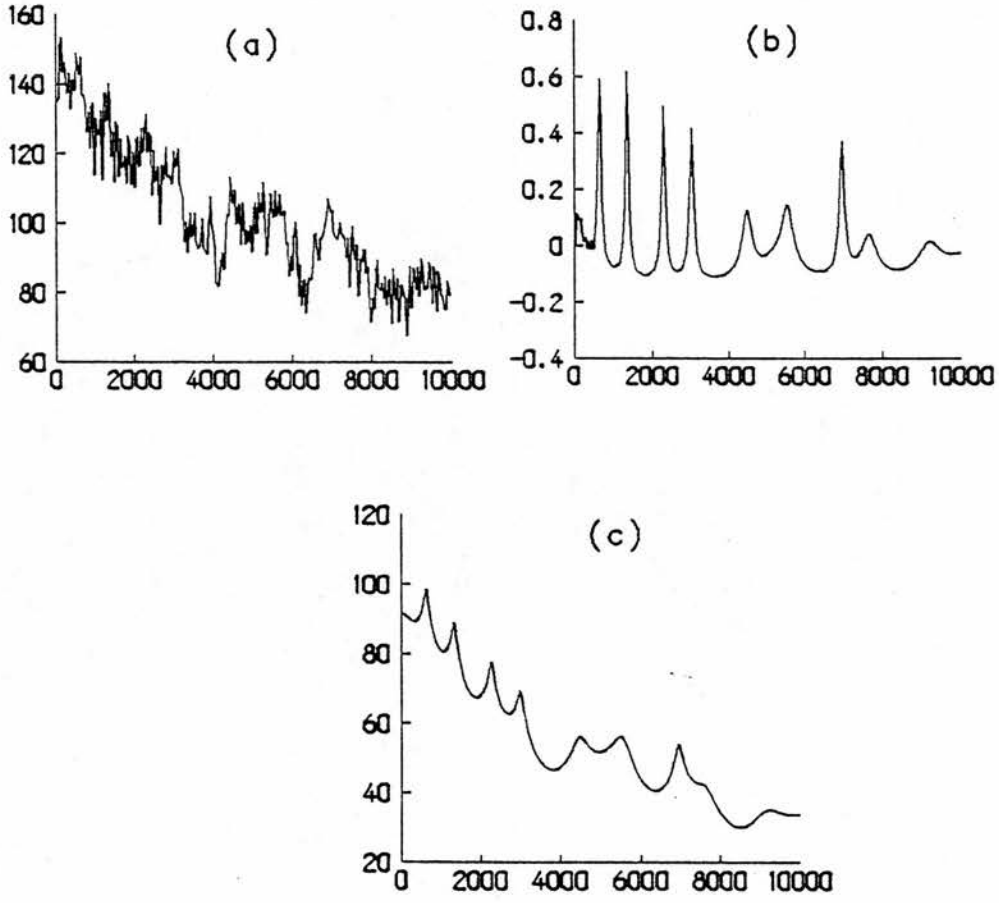

Figure 4.6 Log magnitude spectra for the neutral vowel [@] (a) DFT spectrum (b) Negative Derivative of Phase spectrum (c) Linear Prediction spectrum

representations is clear. Both give a reasonable approximation to the envelope of the log magnitude spectrum produced directly by Fourier analysis (Figure 4.6 (a)). The peaks in the NDPS are seen to correspond to those in the FFT spectrum.

Using this technique, however, only the *peaks* of the spectrum are well represented, because the original analysis, needed to give a smoothed frequency spectrum, is based on the all-pole Linear Prediction model. The pole-zero decomposition method described by Yegnanarayana (1981), and investigated here, builds on his earlier work by extending the use of the NDPS to the estimation of poles *and* zeros. The NDPS is not itself used for formant estimation, however; rather, it is used to achieve a separation of pole and zero activity to allow the derivation of the pole-zero Linear Prediction model.

Yegnanarayana (1981) argues that the properties of the NDPS for zeros will be similar to those of the NDPS for poles, except that it will have a negative sign. Complex conjugate pole pairs in the transfer function of the filter model give rise to positive peaks in the NDPS, while complex conjugate zero pairs in the transfer function give rise to negative peaks in the NDPS. The important property of the NDPS – that individual pole pairs have little influence on each other's contribution to the overall DPS – still holds when zero pairs are included in the transfer function, since the contributions of individual resonators (or anti-resonators in the case of zeros) are additive rather than multiplicative. Therefore it should be possible to separate not only closely spaced formant peaks, but also nearly coincident pole-zero pairs.

The starting point for the calculation of the DPS is the smoothed complex spectrum given by Fourier analysis. In Yegnanarayana (1978), this spectrum is obtained by a DFT on the predictor coefficients obtained by Linear Prediction analysis (12th order). As noted above, the all-pole model means that only the peaks of this smoothed spectrum are well represented. In Yegnanarayana (1981), the smoothed spectrum is obtained by *cepstral* analysis instead, so that the shape of both peaks and dips is maintained for modelling by the pole-zero filter. The negative derivative of phase spectrum is calculated directly from the complex cepstrum, by exploiting a relationship between the two representations. The log spectrum can be derived from the cepstrum by the following relationship:

$$\ln V(\omega) = \tfrac{1}{2}c(0) + \sum_{k=1}^{\infty} c(k) e^{-jk\omega} \tag{4.13}$$

where $\{c(k)\}$ are the cepstral coefficients (Yegnanarayana 1981, Eqn.6). The real and imaginary parts of the log spectrum can therefore be expressed in a way which allows us to derive the phase spectrum from the cepstrum:

$$\ln|V(\omega)| = \tfrac{1}{2}c(0) + \sum_{k=1}^{\infty} c(k)\cos k\omega \tag{4.14}$$

(the real part) and

$$\Theta_V(\omega) + 2\lambda\pi = \sum_{k=1}^{\infty} c(k)\sin k\omega \tag{4.15}$$

(the imaginary part) (Yegnanarayana 1981, Eqn.8,9). Equation 4.15 also represents the *negative phase spectrum*, and the derivative of this (the NDPS) can be obtained by the following relationship:

$$\Theta'_V(\omega) = \sum_{k=1}^{\infty} kc(k)\cos k\omega \qquad (4.16)$$

(Yegnanarayana 1981, Eqn.10). Thus it is possible to calculate the NDPS directly from the cepstrum, without returning to the frequency domain.

The cepstral representation is important because it allows us to smooth the frequency spectrum. This is normally done by selecting only the first few cepstral coefficients for transformation back into the frequency domain (between 20 and 40, at a sampling frequency of 10 kHz: Yegnanarayana 1981: 6). This is illustrated in Figure 4.7, which shows the cepstrally smoothed spectrum for the vowel [@] with the original DFT spectrum for comparison. Note that the cepstrally smoothed spectrum gives equal weight to peaks and dips, even if modelling of the peaks is not as accurate as that given by LPC (Figure 4.6 (c)). In Yegnanarayana (1981), by limiting Equation 10 in a similar manner to only the first M cepstral components, the NDPS of the *smoothed* frequency spectrum is obtained:

$$\Theta'(\omega) = \sum_{k=1}^{M} kc(k)\cos k\omega \qquad (4.17)$$

where M is the model order (Yegnanarayana 1981, Eqn.27).

From this NDPS, it should be possible to calculate formant and anti-formant frequencies and bandwidths, in a similar way to that described in Yegnanarayana (1978): the locations of the peaks (positive or negative) represent the frequency locations of the poles or zeros, while their heights represent their bandwidth.

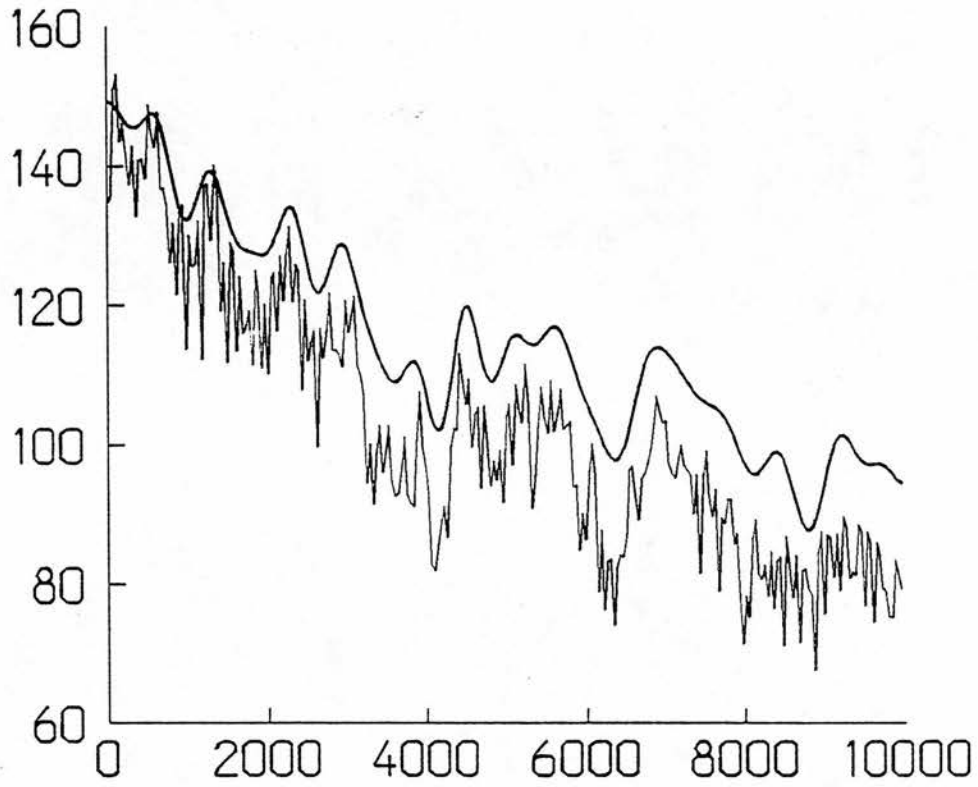

Figure 4.7 DFT magnitude spectrum and cepstrally-smoothed spectrum for [@]

However, since the usual aim of speech analysis is to represent the smoothed spectrum by a mathematical model, Yegnanarayana (1981) uses the NDPS to derive the numerator and denominator coefficients of a pole-zero Linear Prediction model. Since peaks in the positive half of the NDPS reflect pole activity while peaks in the negative half reflect activity of zeros, the positive and negative portions of the NDPS are considered separately in the analysis, to give the denominator and numerator coefficients respectively.

Yegnanarayana (1981) shows that the positive half of the NDPS can be expressed in terms of the cepstral coefficients for the pole-only response:

$$[\Theta'_V(\omega)]^+ = C + \sum_{k=1}^{\infty} kc^+(k)\cos k\omega \qquad (4.18)$$

(Yegnanarayana 1981, Eqn.14), while the negative half can be expressed in terms of the cepstral coefficients for the zero-only response:

$$[\Theta'_V(\omega)]^- = -C + \sum_{k=1}^{\infty} kc^-(k)\cos k\omega \qquad (4.19)$$

where $\{c^+(k)\}$ and $\{c^-(k)\}$ represent the cepstral coefficients for pole and zero spectra of $V(\omega)$ respectively, and C is the average value. (Yegnanarayana 1981, Eqn.15),

This splitting of the *cepstrum* into a pole-only part and a zero-only part is achieved by separating the positive NDPS values from the negative values and returning each half to the cepstral domain separately using the Inverse DFT. (Remember that the NDPS was itself obtained from the cepstrum by the forward DFT.) The resulting cepstral signals represent the smoothed all-pole signal and the smoothed all-zero signal (Figure 4.8). These signals can then be modelled by two independent Linear Prediction analyses. The method used exploits a relationship between the cepstrum and predictor coefficients (Gray and Markel 1976) such that the cepstral coefficients of a stable all-pole filter system can be derived from the predictor coefficients by a simple recursion operation. In Yegnanarayana's method, the reverse of this operation gives the predictor coefficients from the cepstral coefficients. The pole coefficients are calculated from the all-pole cepstral signal

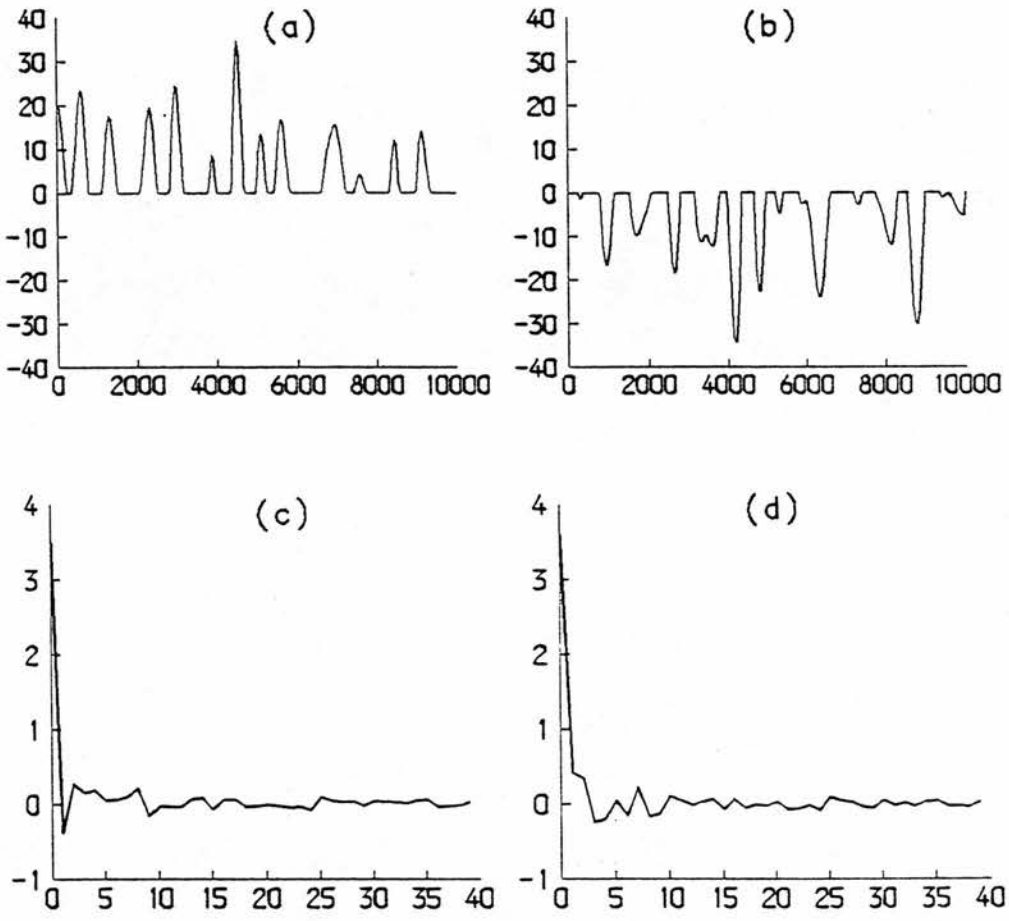

Figure 4.8 Decomposition of the cepstrum via the NDPS (a) positive NDPS values (b) negative NDPS values (c) all-pole cepstral response (d) all-zero cepstral response

$$a^{+}(1) = -c^{+}(1) \tag{4.20}$$

$$ja^{+}(j) = -jc^{+}(j) - \sum_{k=1}^{j-1} kc^{+}(k)a^{+}(j-k)$$

for $j = 2,3,...M$ (Yegnanarayana 1981, Eqn.16), and the zero coefficients from the all-zero cepstral signal

$$a^{-}(1) = c^{-}(1) \tag{4.21}$$

$$ja^{-}(j) = jc^{-}(j) + \sum_{k=1}^{j-1} kc^{-}(k)a^{-}(j-k)$$

for $j = 2,3,...M$ (Yegnanarayana 1981, Eqn.17).

This direct calculation bypasses the minimization problem which is at the heart of other pole-zero modelling techniques. It can, however, be shown that the parameters chosen by this method are those which match the first $M+1$ cepstral coefficients of the model and the first $M+1$ cepstral coefficients of the signal (Yegnanarayana 1981: 10). For this reason, the technique has been referred to as *cepstral matching*, in the same way that Linear Prediction can be viewed as autocorrelation matching.

The resulting pole-zero spectrum is similar to the cepstrally-smoothed spectrum obtained by the usual method, but has greater resolution. This is because the decomposed cepstral responses are not truncated (that is, set to zero beyond M coefficients) as in ordinary cepstral smoothing, but are extrapolated beyond M in the process of modelling by the use of the same number of predictor coefficients as there are cepstral coefficients.

The frequency response of the pole-zero model is obtained by deriving separately the (log) frequency responses of the numerator and denominator polynomials and adding them together. However, considering the two halves separately can aid in the identification of significant zeros or poles which might by obscured in the combined pole-zero spectrum.

Yegnanarayana (1981) notes that there is some interaction between the poles and zeros in the derivative of phase spectrum, owing to the discrete nature of the signal. Thus the splitting of pole and zero contributions is only

approximate. The interaction is more prominent when small numbers of cepstral coefficients are used, and becomes less severe with higher model orders. This is unlike the situation found with other analysis methods, where increasing the model order depletes the information left in the LPC residual for the calculation of the zeros, and where the prior estimation of the *number* of poles and zeros is therefore crucial.

The interaction between the poles and zeros can be seen in the complementary nature of the pole and zero spectra in Figure 4.9. The pole spectrum has narrow peaks and broad valleys where the zero spectrum has broad peaks and narrow valleys. The two combine to give a spectrum which accurately models both peaks and dips in the original short-time spectrum (Figure 4.7).

4.4.1.2. The algorithm

A full description of the algorithm is given in Yegnanarayana (1981). Figure 4.10 illustrates its operation. In the algorithm, the cepstrum is obtained by an inverse DFT on the log power spectrum (calculated from the windowed input speech signal). Spectral smoothing is obtained by selecting the first M cepstral coefficients for use in deriving the NDPS. This is given by weighting these first M coefficients with an increasing ramp function (proportional to k, the index of each cepstral coefficient in Equation 4.17 above) before calculating their Fourier spectrum using the DFT and discarding the imaginary values. This sequence of operations is equivalent to the calculation described in Equation 4.17 above since only the cosine terms are needed.

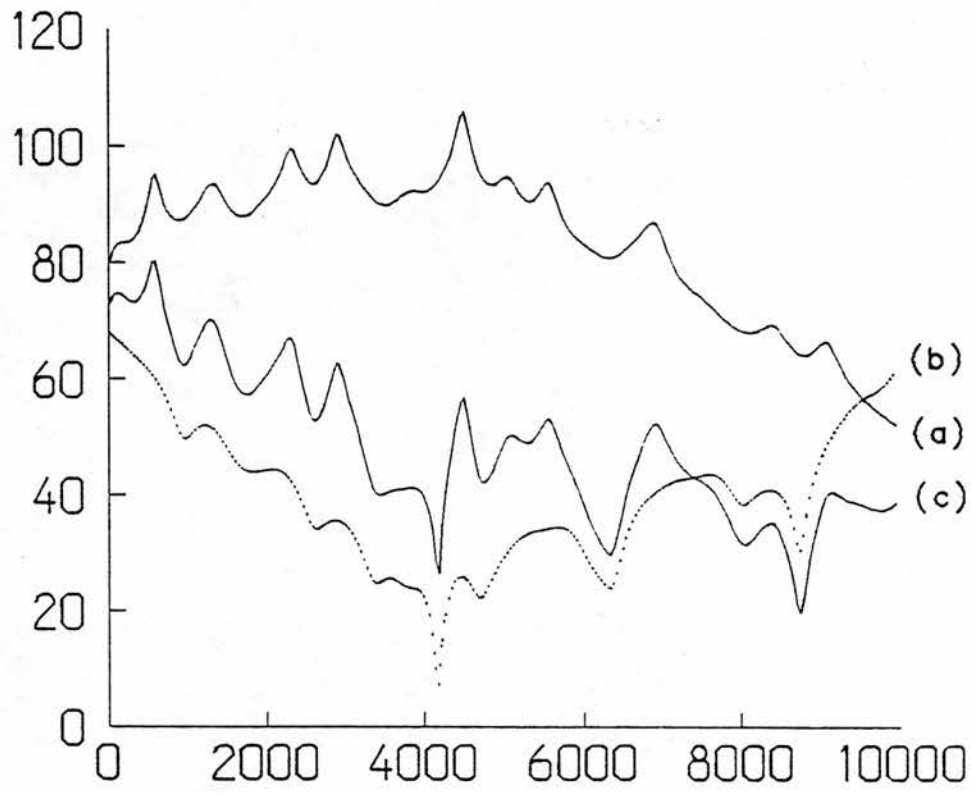

Figure 4.9 Cepstral decomposition spectra for the neutral vowel [@] (a) all-pole response (b) all-zero response (c) pole-zero response

The positive and negative portions of the NDPS are separated by loading them into separate arrays and padding them with zeros where necessary. The positive and negative halves of the NDPS are then returned to the cepstral domain by separate Inverse DFTs, and de-weighted to compensate for the effects of the original weighting. This gives the two cepstral responses $\{C^+(n)\}$ and $\{C^-(n)\}$. The denominator coefficients are calculated from the sequence $\{C^+(n)\}$ by reverse recursion, and similarly the numerator coefficients from the

sequence $\{C^{-}(n)\}$. Equal numbers of numerator and denominator coefficients are obtained: thus the model has equal numbers of poles and zeros. The pole spectrum is calculated from the denominator coefficients by a DFT, and the log magnitude (or power) evaluated; the zero spectrum is calculated from the numerator coefficients, and the log magnitude (or power) evaluated. The arithmetic sum of the (log) pole and zero spectra gives the combined log magnitude (power) frequency response of the pole-zero model.

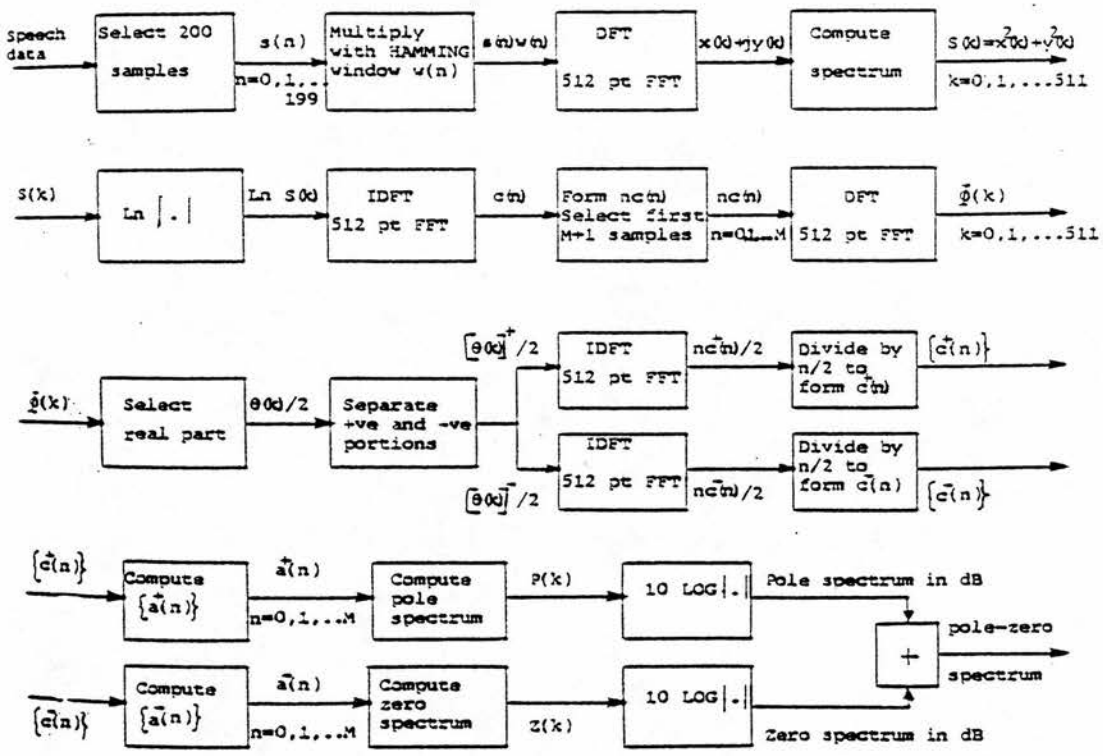

Figure 4.10 The cepstral decomposition algorithm (from Yegnanarayana 1981)

4.5. Testing the pole-zero decomposition method using synthetic signals

4.5.1. Introduction

Yegnanarayana's paper demonstrates that the pole-zero decomposition method gives a good approximation to the short-time spectrum of real speech signals. Interpretation of such spectra can be a problem, however, because it is not possible to identify all peaks with poles and all dips with zeros. Some peaks result from the closeness of adjacent zeros in the spectrum. Figure 4.11, showing the pole-zero spectrum (and LPC spectrum) of a voiced fricative, gives an example of this: the closeness of zeros at 2.3 kHz and 2.7 kHz produces a sharp peak at 2.5 kHz in the combined frequency response.

For this study, it would be useful to be able to identify pole and zero activity separately, to assess its potential for speaker verification. It was decided, therefore, to test the performance of the pole-zero modelling technique on a synthetic signal whose pole-zero distribution was known. This would aid interpretation of the separate pole and zero responses provided by the model, and help to assess the degree of interaction between them. Of particular interest was the response of the pole-zero system to signals generated by an *all-pole* or *all-zero* digital filter: this would aid interpretation of the significance of the zero response provided by the model for the velar nasal stop /ng/, for which the presence of zeros is a matter still to be resolved. In addition, the ability of the system to separate known pole and zero contributions to an observed pole-zero signal was investigated. A synthetic signal was used

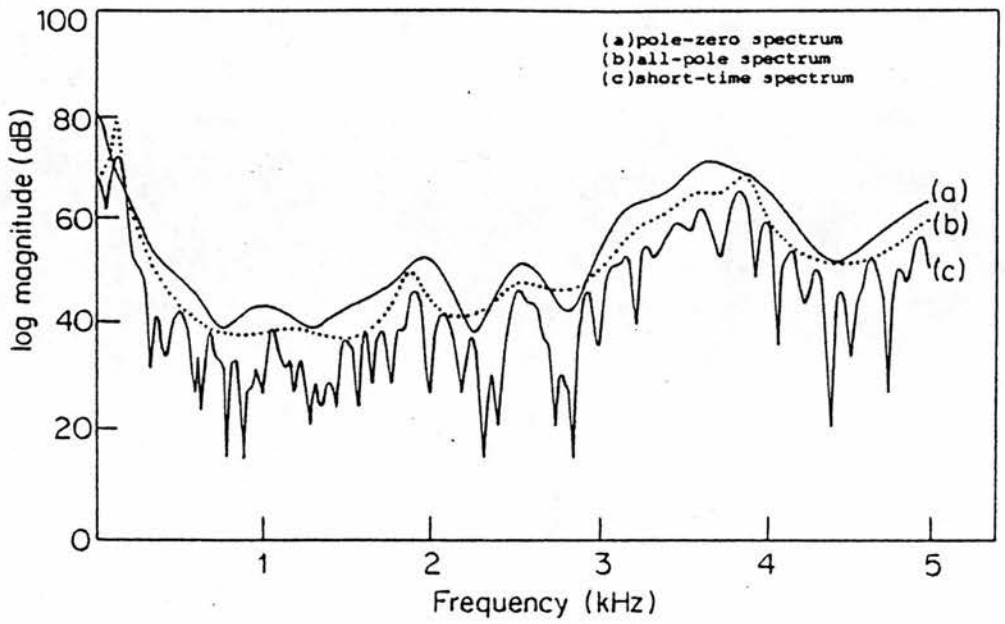

Figure 4.11 Spectra of a voiced fricative (from Yegnanarayana 1981) (a) pole-zero spectrum (M=20) (b) LP all-pole spectrum (M=20) (c) FFT spectrum

because it was not possible to be sure of the pole-zero composition of a real speech sample.

4.5.2. Outline of experiment

Synthetic signals were generated by a software implementation of a digital filter, excited by a single impulse. The resulting time signal was treated as a speech waveform sampled at 10 kHz, and modelled using an implementation of the Yegnanarayana technique programmed by Mr Clive Summerfield (Summerfield 1988). Pole response, zero response and combined pole-zero response spectra were derived and compared with the DFT-derived impulse

frequency response and a Linear Prediction smoothed spectrum.

4.5.3. Design of digital filter

The digital filter was a software implementation written in C of the general pole-zero discrete-time filter illustrated by Bozic (1979, Fig.2.4), shown here in Figure 4.12. The filter operates on each input (excitation) value in turn to produce an output value which is the weighted sum of past output values, past input values and the present input value. Input values are weighted (multiplied) by the coefficients of the numerator (a_0 to a_N in Figure 4.12), and the output values weighted by the coefficients of the denominator (b_1 to b_M) before being combined by addition. The numerator coefficients represent the effect of the zeros of the filter's transfer function, the denominator coefficients the effect of its poles. The filter's design is thus identical to that of the general pole-zero model used in Linear Prediction (see 4.2), the coefficients a_1 to a_N, b_1 to b_M being the predictor coefficients. The coefficient a_0 is simply a scaling factor, here set to 1.0 throughout.

The operation of the filter can be illustrated as follows. At, for example, the fourth sample in the input, $k=3$, the output sample $y(3)$ equals the sum of the weighted (present and past) inputs minus the sum of the weighted (preceding) outputs:

$$y(3) = (x(3)*a(0) + x(2)*a(1) + x(1)*a(2) + x(0)*a(3)) \qquad (4.22)$$
$$- (y(2)*b(1) + y(1)*b(2) + y(0)*b(3))$$

The first output sample, $y(0)$, is simply equal to the first input sample, $x(0)$, times the scaling factor $a(0)$.

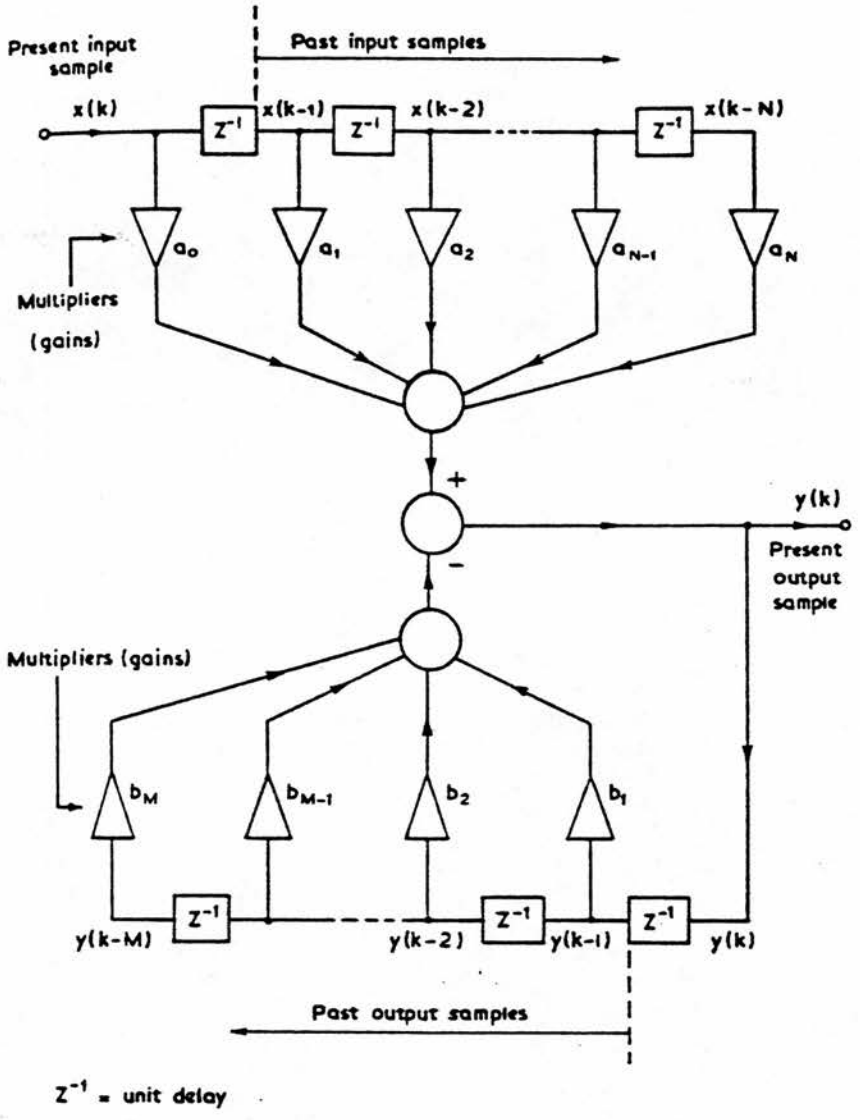

Figure 4.12 Implementation of a general pole-zero discrete time filter (from Bozic 1979)

The frequency response of the filter is given by combining the frequency responses of the numerator and denominator polynomials:

$$H(z)=\frac{N(z)}{D(z)} \tag{4.23}$$

These are easily calculated by a DFT on the coefficients themselves (see, for

example, Figure 4.13 below). Alternatively, the filter can be excited by an impulse or white noise to give a time-based signal (the impulse response: see Figure 4.14 below); a DFT of this impulse response gives the (impulse) frequency response (see Figure 4.15 below). It is generally simpler to derive the frequency response directly from the coefficients, as is done below. To ensure comparability between the pole-zero, LPC and DFT spectra, however, the DFT of the impulse response of each filter was also derived.

4.5.4. Choice of coefficient values

Coefficients were chosen to give suitable all-pole, all-zero and pole-zero signals for input to the pole-zero decomposition analysis. The choice of pole and zero coefficients was made simply by plotting the log power spectrum of the DFT of different sets of coefficients, arbitrarily chosen, until a reasonable approximation to an imaginary speech spectrum was obtained. The "sampling frequency" of these digital signals was set to 10 kHz: that is, it was assumed that the points in the input (and output) sequences represented samples taken from a waveform at intervals of 0.0001 seconds. The resulting power spectra therefore represented information up to 5 kHz (the Nyquist frequency). In each case, the filter coefficients were zero-padded and a 256 point DFT was computed, giving 128 frequency bins spanning 5 kHz.

All-pole filter ($M=10$)

A tenth-order all-pole filter was obtained by setting all the numerator coefficients a_1 to a_M equal to zero. The scaling factor a_0 remained at 1.0 to

ensure that the filter produced an output for the denominator coefficients to process. Ten denominator coefficients b_1 to b_{10} were found as described above, by choosing values arbitrarily, plotting their DFT log power spectrum and adjusting them until a suitable spectrum was obtained. The log power spectrum of the ten coefficients chosen was *inverted* to give the frequency response of the all-pole filter; this is shown in Figure 4.13 (a). The values of the denominator coefficients were as follows:

b_1 0.75
b_2 0.3
b_3 0.2
b_4 0.5
b_5 0.4
b_6 0.2
b_7 0.6
b_8 0.3
b_9 0.2
b_{10} 0.05

All-zero filter ($M=2$)

A second-order all-zero filter was obtained by setting all the denominator coefficients b_1 to b_{10} to zero and selecting two numerator coefficients a_1, a_2 in the way described above. The scaling factor a_0 remained at 1.0. The log power spectrum of the resulting all-zero filter is shown in Figure 4.13 (b). The values of the coefficients were:

a_1 0.9
a_2 0.6

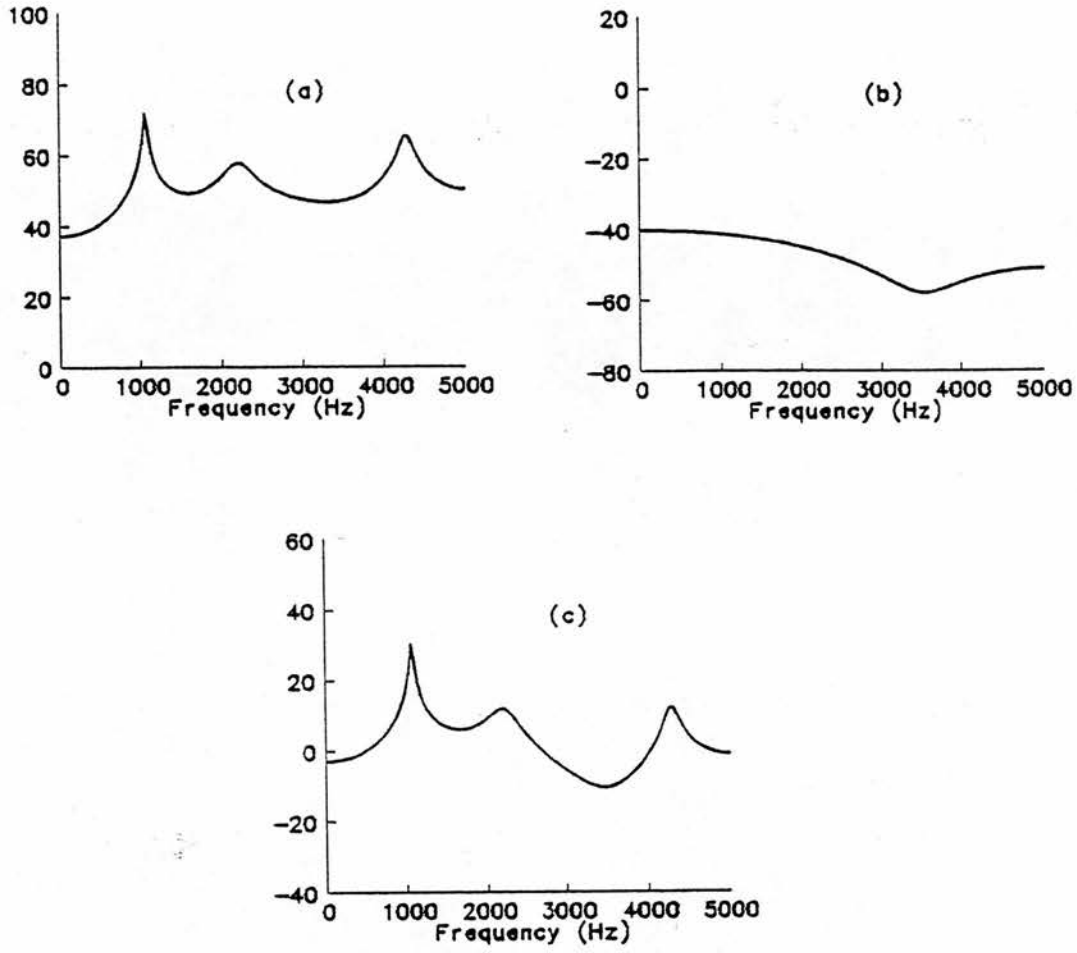

Figure 4.13 Frequency response spectra of (a) all-pole filter (M=10) (b) all-zero filter (M=2) (c) pole-zero filter

Pole-zero filter (M=10,2)

A pole-zero filter was obtained by using the ten denominator (pole) coefficients from the all-pole model and the two numerator (zero) coefficients from the all-zero model. The frequency response of the combined filter was obtained by dividing the zero response by the pole response (or equivalently, by subtraction of their log power spectra). The combined frequency response is shown in Figure 4.13 (c).

4.5.5. Pole-zero decomposition of synthetic signals

4.5.5.1. Generation of the synthetic signals

The filters whose frequency responses are presented above were then excited by an impulse to produce a time-based signal for analysis by the pole-zero decomposition method. The excitation consisted of a 256-point sequence:

0 2000 0 0 0

The sequence was made to begin with 0, rather than with the impulse sample, so that the output signal began and ended with 0; this removed the need for any windowing of the input signal during the DFT. The value of the impulse was chosen to be 2000 in order to give a reasonable range of output values, between -2048 and 2047 (the range achievable by 12-bit quantization of a digitized speech waveform). The impulse response output signals for the three filters (all-pole, all-zero and pole-zero) are shown in Figures 4.14 (a,b,c). Each point in the resulting waveform is assumed to be a sample from a signal digitized at 10 kHz, giving an interval of 0.0001 secs between points; thus these waveforms each represent 25.6 milliseconds.

Each waveform was analyzed using the cepstral decomposition technique described above. However, the input samples were not windowed, since the sequences for analysis all began and ended with 0; thus the assumption of an infinitely repeating waveform made by the DFT was not violated. Pre-emphasis was not applied, since this would have introduced a zero not present in the input. Each waveform was analyzed as a single 25.6 millisecond frame, using an FFT size of 256 points (2^8). Pole-zero models of order $M=10$, $M=40$

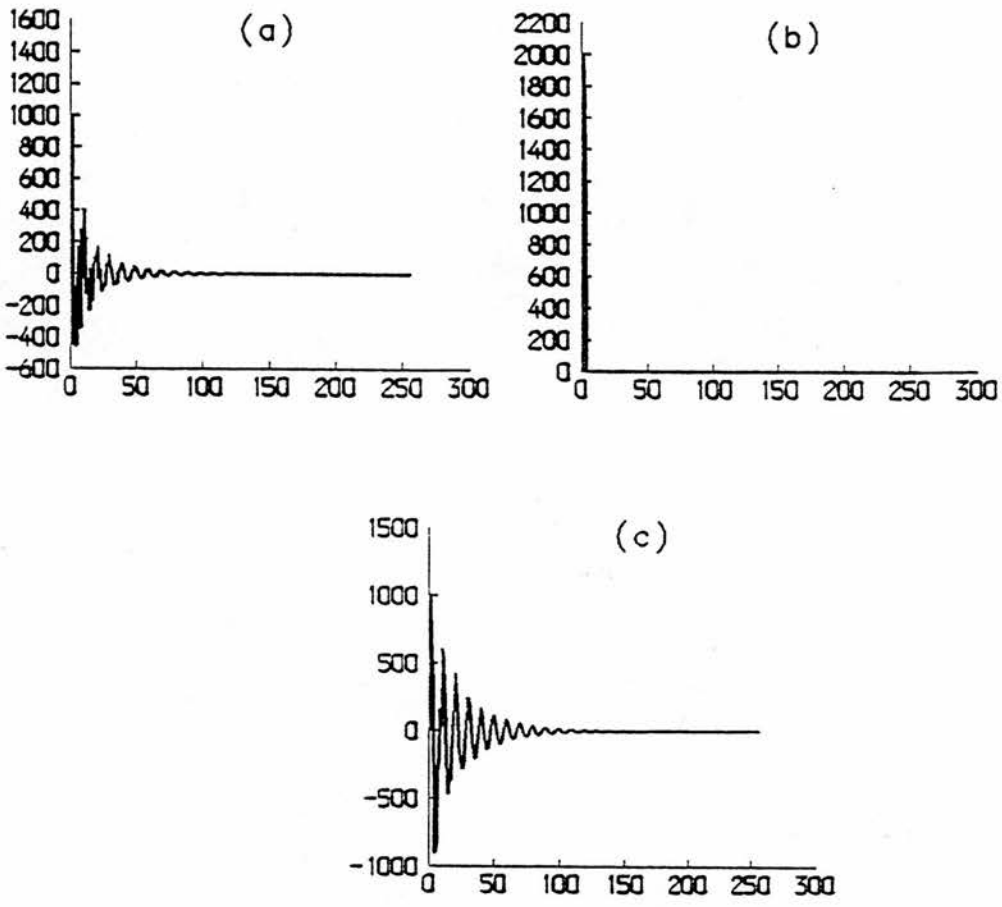

Figure 4.14 Impulse response of (a) the all-pole (b) the all-zero and (c) the pole-zero filters

and $M=60$ were derived for each waveform in separate analyses, to study the effects of increasing the model order. Since only a single impulse was used to excite the filter, there was no danger of the higher model orders causing the output spectra to reflect the periodicity normally found in speech waveforms; however, these model orders are probably higher than could be attempted with real speech (at least for females), since at a sampling rate of 10 kHz the first 60 cepstral coefficients would include the excitation information (the "spike"

caused by the periodicity) for many female speakers.

A Linear Prediction analysis of each waveform was also carried out. To allow comparison with the pole-zero analysis, no pre-emphasis or windowing were applied. Neither was strictly necessary with these signals, and analyses *with* pre-emphasis ($\mu = 0.95$) and a hanning window (not presented here) gave similar spectra, but slightly smoother. All three signals (all-pole, all-zero and pole-zero) were analysed using a 24th-order predictor. Log magnitude spectra were derived from the predictor coefficients via a 512-point DFT. Figure 4.15 shows the spectra for the 24th-order analysis. It can be seen that, while the response to the all-pole signal (Figure 4.15 (a)) is adequate, the response to the all-zero signal (Figure 4.15 (b)) completely fails to model the behaviour seen in the DFT spectrum (Figure 4.13 (b)). Similarly, the LPC model of the pole-zero signal (Figure 4.15 (c)) shows reasonably accurate peaks, but inaccurate modelling of the major dip. Figure 4.15 (d) shows a 40th-order Linear Prediction analysis for the pole-zero signal alone: the model's representation of the spectral dip caused by the zeros has improved, but it is still poorly defined, and a ripple has been introduced to the rest of the spectrum.

4.5.5.2. All-pole signal

Figure 4.16 shows the log power spectra (all-pole response, all-zero response and pole-zero response) generated by the analysis for the all-pole signal whose frequency response was given in Figure 4.13 (a) above. A model order of $M = 10$ (Figure 4.16 (a)) gives a rather smooth pole response and a zero response which is complementary in having peaks and dips corresponding to

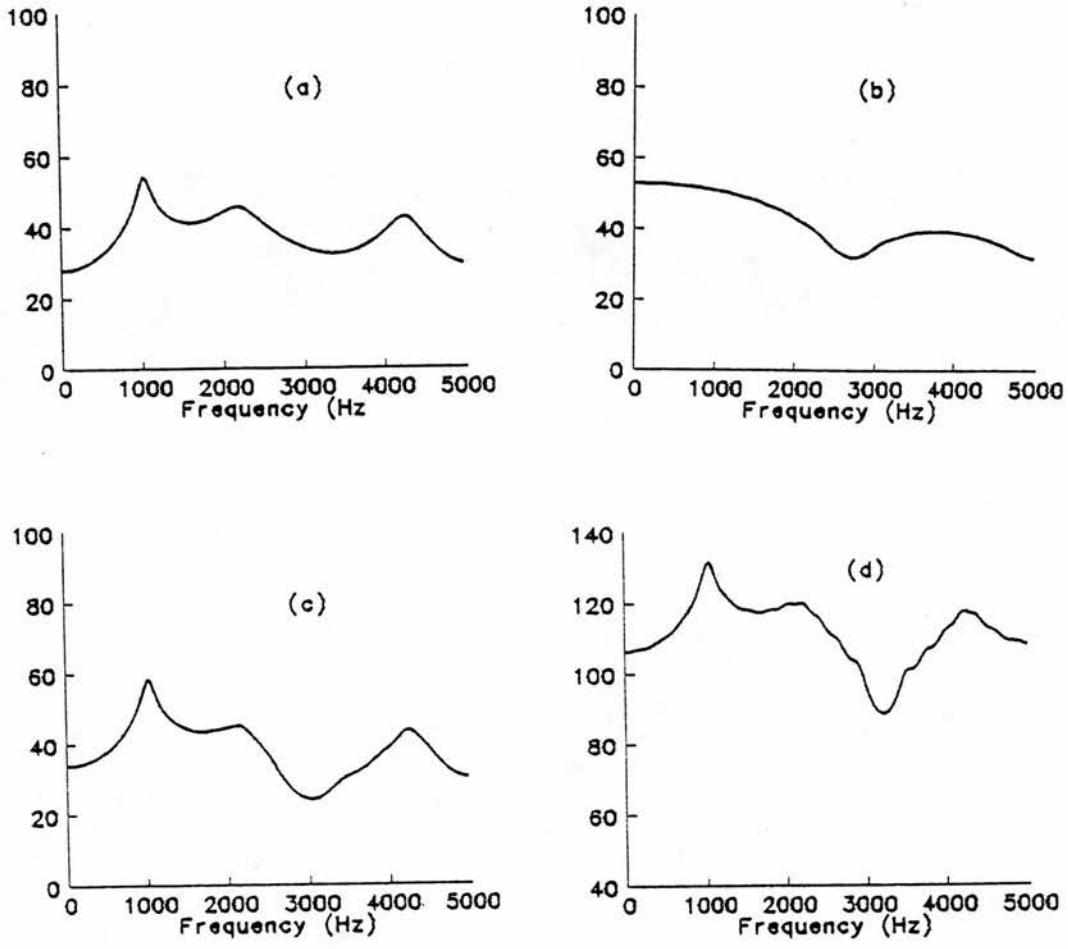

Figure 4.15 Linear Prediction spectra of synthetic signals (M = 24) (a) all-pole signal (b) all-zero signal (c) pole-zero signal (d) pole-zero signal, M = 40

those in the pole response. This highlights one problem with interpreting the separate pole and zero frequency responses, especially at low modelling orders: even though no zeros were present in the input signal, the zero response shows some spectral structure.

However, increasing the model order to $M = 40$ gives a better analysis: the peaks of the pole response have become better defined, while the zero response has become *flatter* (i.e. less pronounced). Very similar results are obtained with

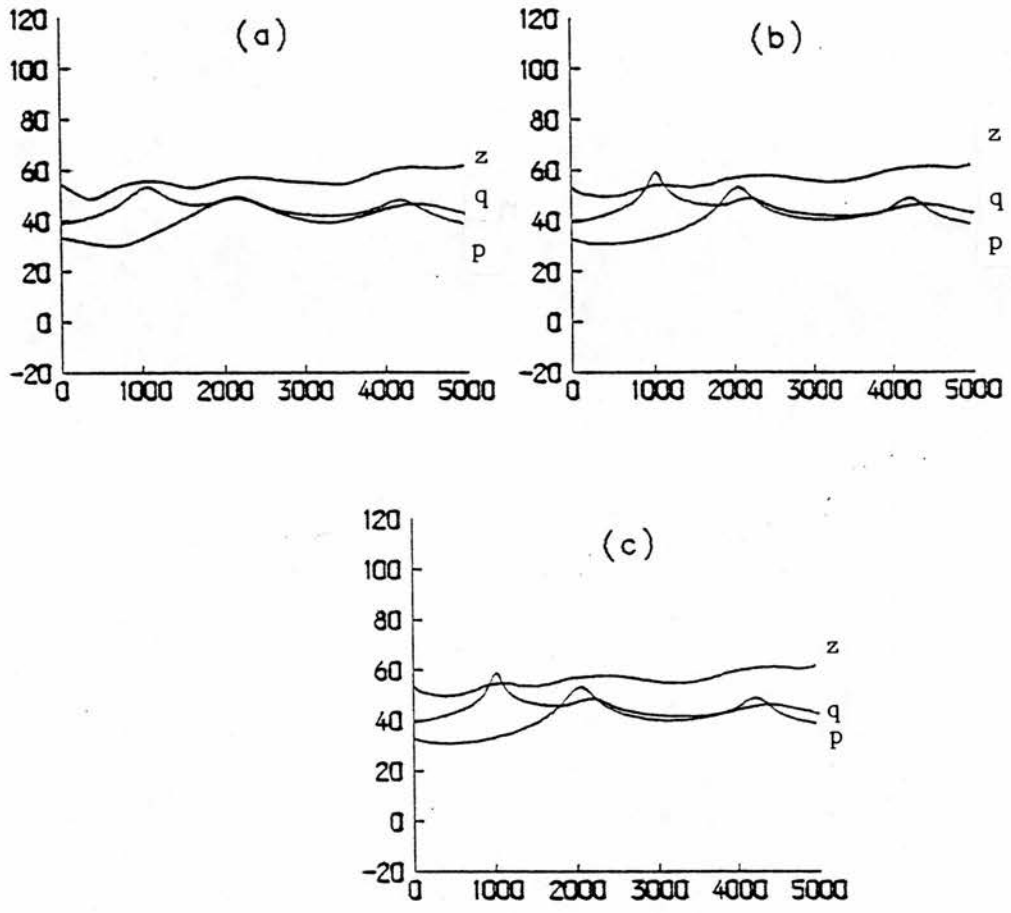

Figure 4.16 Pole-zero spectra derived for the all-pole signal with increasing model order; p = all-pole, z = all-zero, q = pole-zero response

a model order of $M=60$, so over-estimation of the number of poles and zeros required seems to have had no detrimental effects.

4.5.5.3. All-zero signal

Figure 4.17 shows the log power spectrum (all-pole response, all-zero response and pole-zero response) generated by the analysis for the all-zero signal whose frequency response was given in Figure 4.13 (b) above. A model order of $M=10$ (Figure 4.17 (a)) gives both a zero spectrum and a pole-zero

spectrum which are good approximations to the designed frequency response. The all-pole response shows some correlation with the high and low points of the all-zero response, but no detail. Increasing the model order to $M=40$ (Figure 4.17 (b)) and to $M=60$ (Figure 4.17 (c)) has no noticeable effect on the model's output. The lowest model order ($M=10$) is clearly adequate for such a low order filter (M was equal to 2). Again, gross overestimation of the model order required has had no detrimental effects on the accuracy of the model.

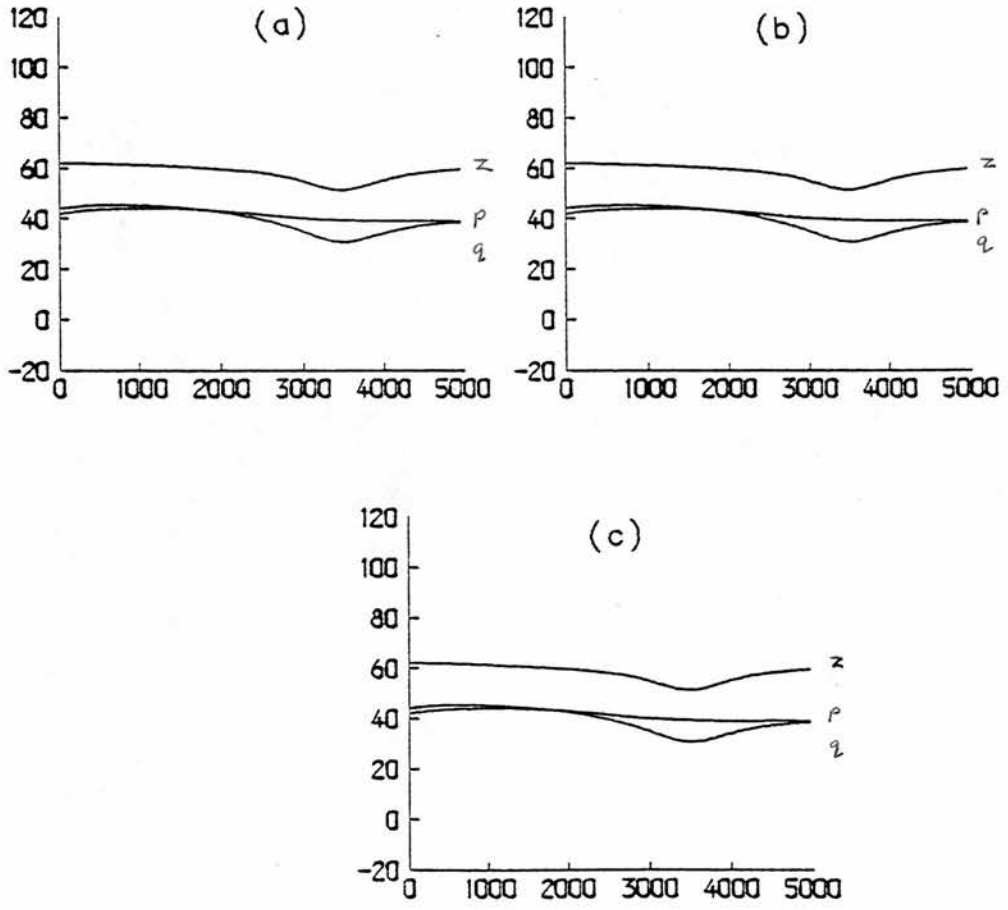

Figure 4.17 Pole-zero spectra derived for the all-zero signal with increasing model order; p = all-pole, z = all-zero, q = pole-zero response

4.5.5.4. Pole-zero signal

Figure 4.18 shows the log power spectra (all-pole, all-zero and pole-zero responses) for the pole-zero signal whose frequency response was illustrated in Figure 4.13 (c) above. A model order of $M=10$ gives an inaccurate representation of both the all-pole response and the all-zero response. The interaction between the pole and zero responses at this low model order can be seen particularly in the zero response, which now shows poor modelling of the major dip at 3000 Hz, a spurious second dip around 500 Hz, corresponding to a low point in the pole response spectrum, and two very broad (but still discernible) features corresponding to the first two peaks of the all-pole response.

Increasing the model order to $M=40$ improves the accuracy of the model. Both the pole and zero responses are more accurate: the zero spectrum has less detail in the (spurious) dip at 500 Hz, and a more pronounced dip at 3000 Hz corresponding to that in the designed frequency response. No great improvement is obtained by a further increase of model order to $M=60$.

4.5.6. Discussion

The response of the pole-zero modelling technique to these signals suggests that it is a highly effective method of modelling, especially where the number of poles and zeros in the transfer function which is being modelled cannot be estimated in advance. Rather high model orders must be used compared with those typical of all-pole modelling (25 coefficients for speech sampled at 20 kHz, according to Markel 1971b), but over-estimation of the model order does not alter the shape of the estimated response. However, there will be an upper

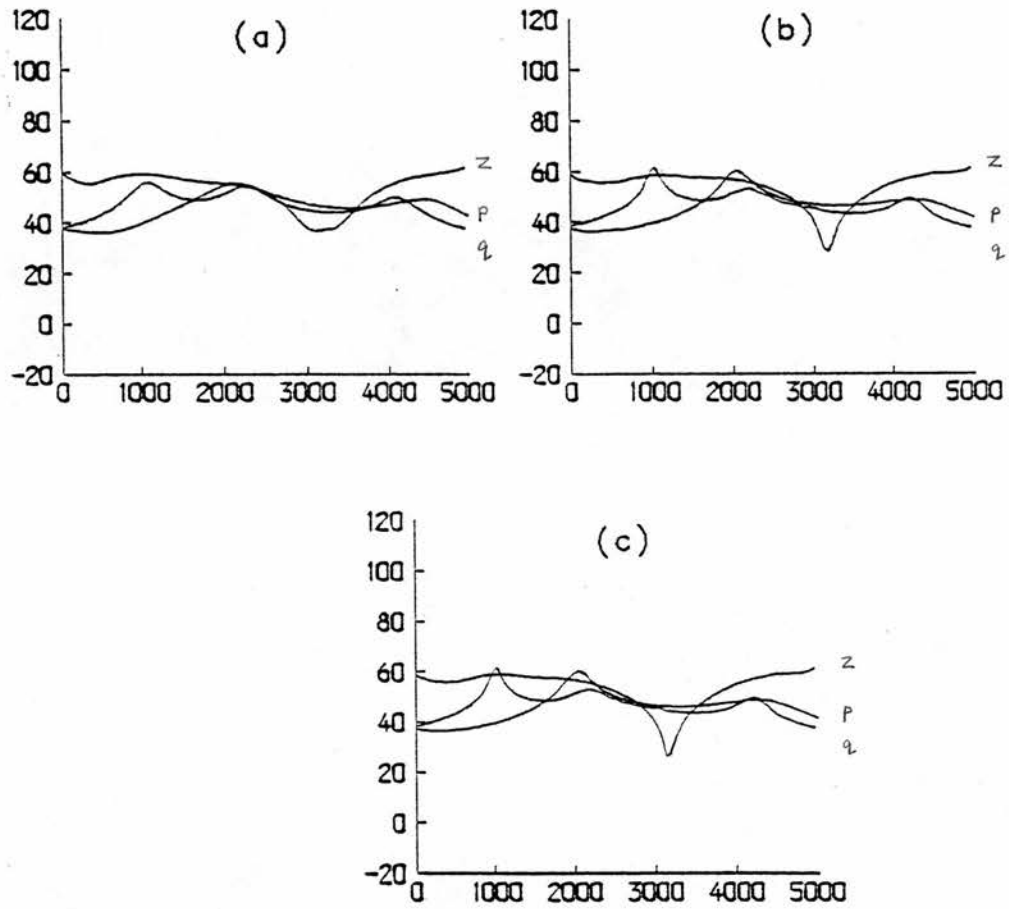

Figure 4.18 Pole-zero spectra derived for the pole-zero signal with increasing model order; p = all-pole, z = all-zero, q = pole-zero response

limit on the model order, set by the minimum pitch period expected in the speech to be processed. Even at high model orders ($M=60$), however, complete separation of the pole and zero components in the signal is not achieved: both the all-pole response to the all-zero signal (Figure 4.17 (c)) and the all-zero response to the all-pole signal (Figure 4.16 (c)) show slight activity. In addition, the all-pole response to the pole-zero signal (Figure 4.18) is not identical to the all-pole response to all-pole signal (Figure 4.16), which is slightly flatter;

however, the frequencies of the peaks of the spectrum match very closely.

Examination of the output spectra in this experiment also suggests that formant estimation from the separate pole and zero frequency responses should be as reliable as that from the combined model response. Some form of peak and dip selection must be used, however, to eliminate features which stem from the unavoidable interaction between the pole and zero models.

4.6. Summary

The use of a pole-zero model for describing signals containing both poles and zeros is widely regarded as desirable, but not generally attempted. In this chapter, an implementation of the modelling method proposed by Yegnanarayana (1981) has been shown to give a good approximation to the spectrum of a synthetic signal containing both poles and zeros. The method appears to be preferable, on theoretical grounds, to other methods proposed in the literature, at least for the study of nasality. In the next chapter, the pole-zero decomposition method will be applied to nasal tokens from real speech, in a more stringent test of its capabilities.

CHAPTER FIVE

CEPSTRAL DECOMPOSITION APPLIED TO NASAL STOPS

CHAPTER FIVE

CEPSTRAL DECOMPOSITION APPLIED TO NASAL STOPS

5.1. Introduction

In Chapter 4 it was concluded – on the basis of experiments with noise-free synthetic signals whose composition was known – that the pole-zero decomposition method was suitable for obtaining separate all-pole and all-zero responses from speech. In this chapter the experiments are extended to use real speech tokens (nasal segments spoken in isolation) whose exact composition is unknown and which were recorded in a rather noisy environment. After some initial experiments in which the limitations of the technique are explored, analysis results for three nasal stops – bilabial, alveolar and velar – are compared with some published data. Some recommendations are made on the use of this technique with real speech.

5.2. Limitations on the pole-zero model order: two experiments

The analysis presented in the last chapter suggested that an increase in model order reduced the interdependence of the separate pole and zero models, though some interaction remained even at high model orders. The use of real speech imposes limitations on how high the model order can go, however. The model order determines how many cepstral coefficients are selected for the

derivation of the smoothed Negative Derivative of Phase Spectrum (see 4.4.1.2). There is an upper limit on this number set by the location of the peak in the cepstrum corresponding to the fundamental periodicity of the speech waveform: if the cepstral coefficients corresponding to this peak are used, the resulting smoothed spectrum will show the harmonic ripple seen in the original power spectrum (Childers et al. 1977). In other words, the separation of *source* and *filter* characteristics will have been lost.

Some compromise must be found so that the response of the model remains faithful to the original spectrum, without introducing frequency ripple or sacrificing the potential of the technique for independent pole and zero estimation.

Normal practice in spectral smoothing by cepstral analysis is to select the cepstral coefficients below the minimum expected pitch period length (Schafer and Rabiner 1970), though an alternative might be to vary the number of coefficients dynamically as the measured fundamental frequency changed. Obviously, the minimum expected pitch period will vary from speaker to speaker, and from male to female subjects. An average figure for male speakers is somewhere between 8 and 10 milliseconds (corresponding to a fundamental frequency of between 125 and 100 Hz); for females this figure will be lower, perhaps down to 4 milliseconds (250 Hz). It should therefore be possible, in theory at least, to use a higher model order for male speakers than for females (for a given speech sampling rate), without running the risk of including pitch period information. As Childers and co-workers (1977: 1438) point out, how-

ever, vocal tract information (convolved with the spectrum of the glottal source) is generally confined to the first 5 milliseconds of the cepstrum, so using coefficients above this limit is unlikely to be of much use.

5.2.1. Outline of the experiments

Two experiments were carried out to choose an appropriate pole-zero analysis model order within the constraints imposed by the need to avoid pitch information and the need to separate the pole and zero contributions. Both were based on pole-zero analysis of single, isolated tokens of the bilabial nasal [m], the alveolar nasal [n] and the velar nasal [ng], recorded by a male speaker of British English (the author). In all cases power spectra for each analysis were averaged over the duration of the token.

In the first experiment, the three nasal tokens were analysed using a range of pole-zero model orders, and the pole-zero response spectra were compared with the FFT spectrum to gauge their accuracy. In the second experiment, the effect of reducing the model order for the zero component of the model was explored using only the bilabial nasal.

5.2.2. Speech tokens

One sustained utterance of each of the three nasal stops of English (bilabial [m], alveolar [n] and velar [ng]), each preceded by the neutral vowel [@], was recorded directly on to a Masscomp MC550 minicomputer, via a Sony PCM-F1 digitiser, an (analogue) anti-aliasing filter set to cut off at 7.5 kHz, and the Masscomp 12-bit analogue-to-digital (A-D) converter. The recording

was made in a computer laboratory, with a moderate level of background noise from machines. The speech was sampled initially at 20000 samples per second, but was later downsampled as described below.

From the middle of each recorded nasal token, a 200 millisecond portion was extracted using a waveform editing tool. Each waveform was then digitally low-pass filtered to 5 kHz using a 256 point rectangular filter, and downsampled at 10 kHz. The downsampling was done because it was expected that there would be minimal formant information in the signal above 5 kHz, and that noise would predominate above this frequency. Because the derivation of the cepstrum involves a non-linear logarithmic operation, this noise – though perhaps of low power – would be likely to interfere with the regions of interest in the cepstrum far more than it would in ordinary spectral analysis (Childers et al. 1977: 1435). The lower sampling rate also reduced the processing required by allowing a shorter Fourier transform, a reduced number of cepstral coefficients and a smaller pole-zero model order.

5.2.3. Experiment 1: the effect of increasing the pole-zero model order

5.2.3.1. Acoustic analysis

Each of the three downsampled tokens was submitted to both an FFT and a pole-zero analysis. The waveform was divided into non-overlapping 25.6 millisecond (256 point) frames; each was multiplied by a Hanning window to reduce discontinuity effects, and a pre-emphasis filter ($\mu = 0.97$) was applied to compensate for the low-pass filtering effects of the vocal tract (Witten 1982:

65-66). The log power spectrum was obtained using a 256 point FFT analysis and averaged over all frames in a given token. A 256 point pole-zero analysis was applied to generate pole-zero models of order 10, 20, 30 40 and 50 coefficients. The largest model corresponded to a cepstrum length of 5 milliseconds (at the sampling frequency of 10 kHz), the upper limit of the vocal tract information in the cepstrum (and well into the region in which pitch-related peaks might be found for female speakers). A model order of 10 required 10 cepstral coefficients and gave a 10th-order all-pole model *and* a 10th-order all-zero model. For each model order, the log frequency responses of the all-pole and all-zero components were derived, and summed to give the response of the combined pole-zero model. Again, the response spectra were averaged over all frames in a token.

5.2.3.2. Effects of increasing model order

Figures 5.1 to 5.3 show the FFT spectrum and the all-pole, all-zero and pole-zero model frequency responses for [m], [n] and [ng] respectively for frequencies up to 5 kHz.

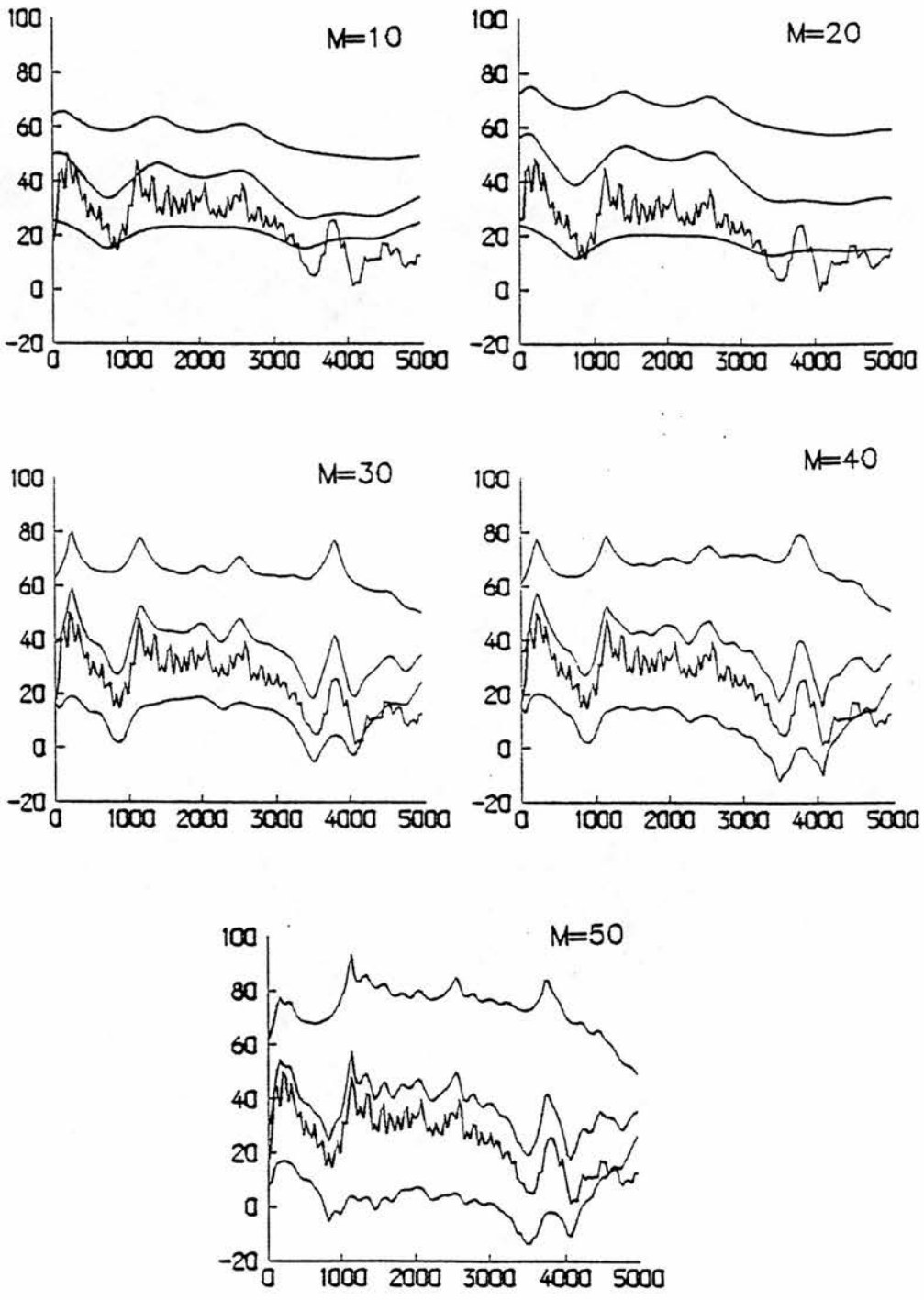

Figure 5.1 All-pole, all-zero and pole-zero spectra for [m] with increasing model order (FFT spectrum shown for comparison)

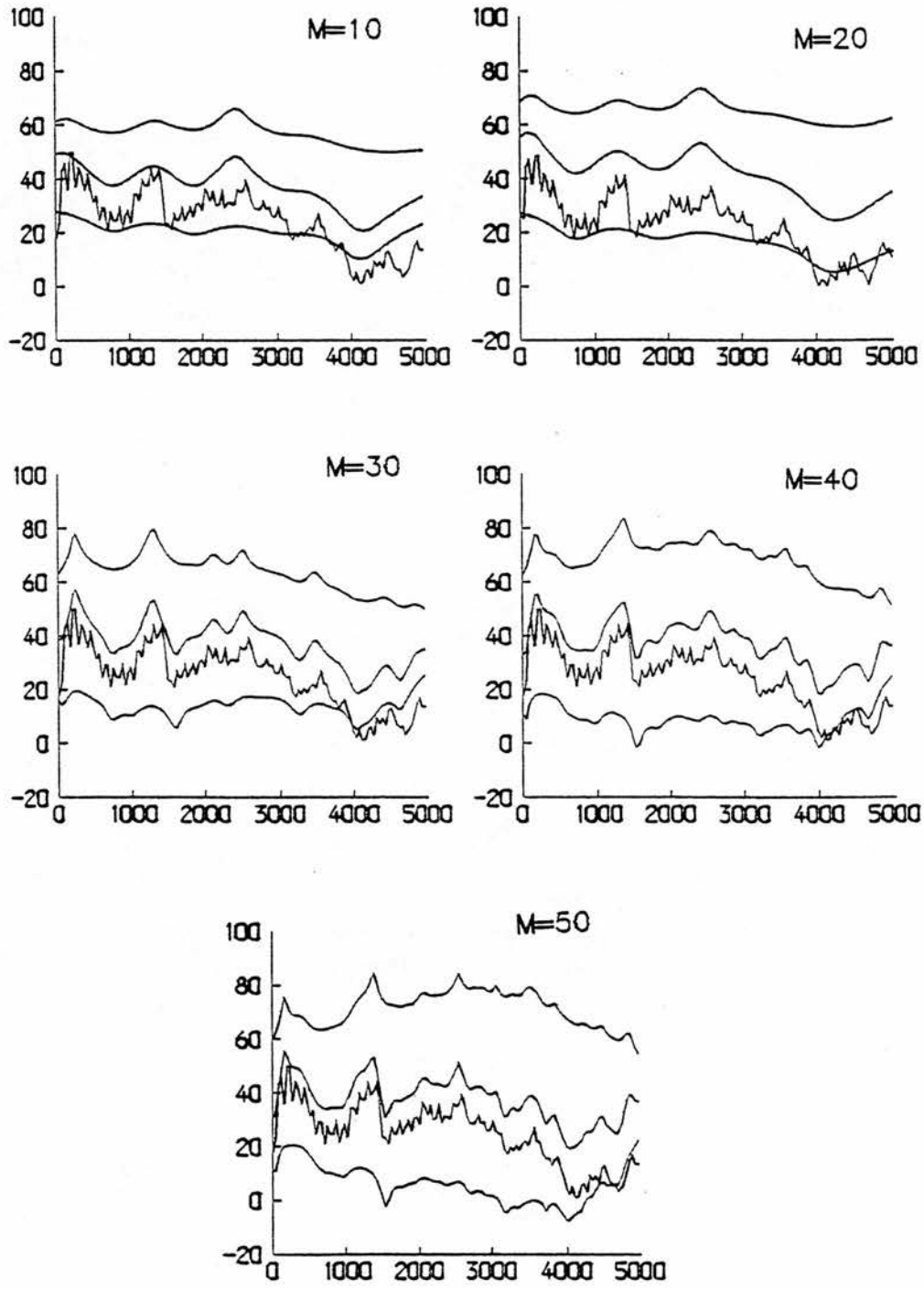

Figure 5.2 All-pole, all-zero and pole-zero spectra for [n] with increasing model order (FFT spectrum shown for comparison)

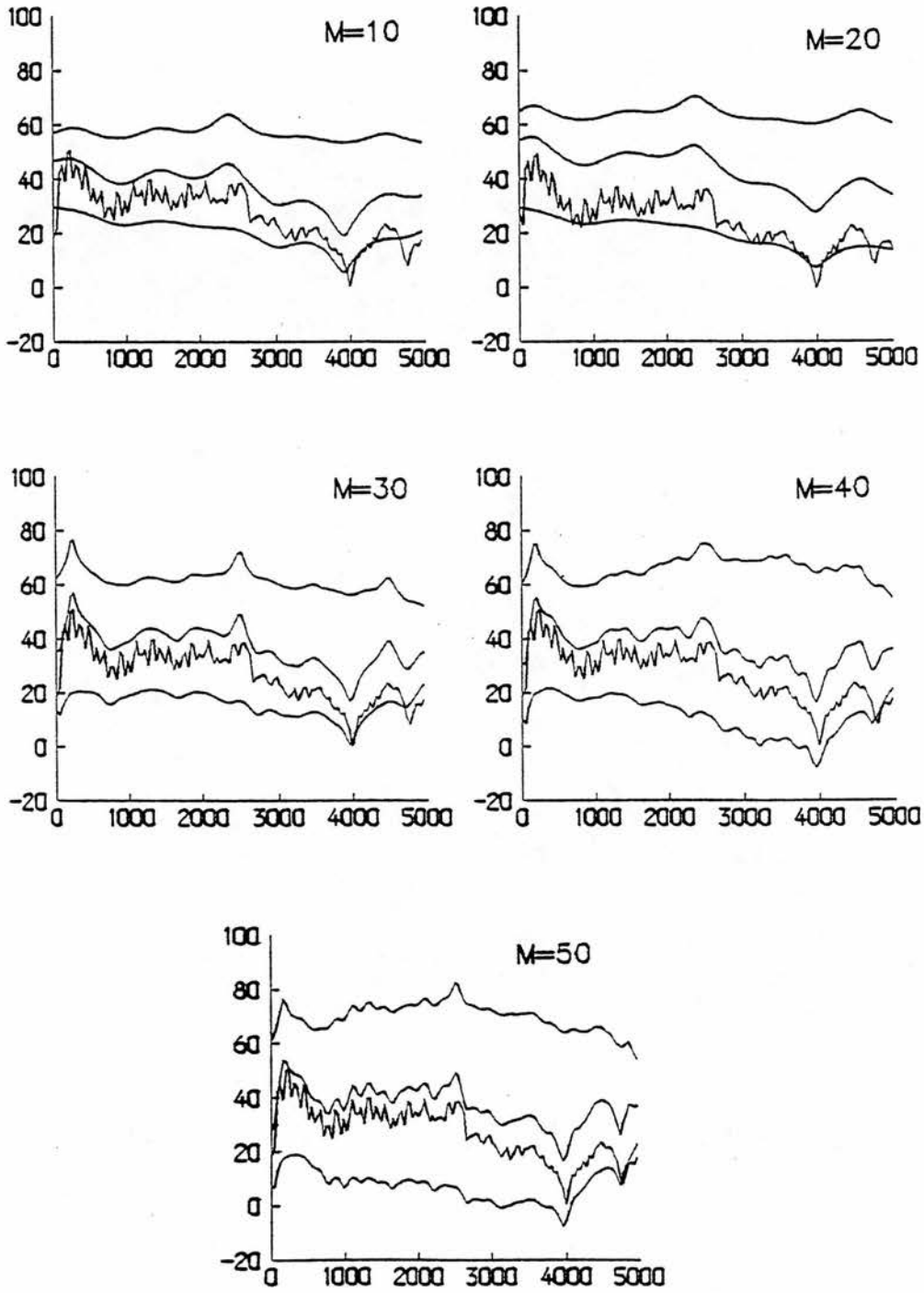

Figure 5.3 All-pole, all-zero and pole-zero spectra for [ng] with increasing model order (FFT spectrum shown for comparison)

In general, the fit of the *pole-zero* spectrum to the FFT spectrum improves with increasing model order. A model order of 10 gives a poor approximation to the FFT spectrum, while an order 30 gives a reasonably faithful result (much more so than order 20). At an order of 50, the pole-zero spectrum becomes rather rough, and formant peaks are difficult to locate. However, the improvement in the combined pole-zero response brings with it two problems: a pronounced bowing of the separate all-pole and all-zero responses (positive in the former, negative in the latter), and a deterioration in the all-zero model, which becomes very noisy by order 30 (especially in the case of the alveolar and velar nasals).

The bowing of the pole and zero spectra is cancelled out in the combined response, but makes estimation of formant frequencies from the separate responses rather difficult. Its cause is not known. The deterioration in the all-zero model response cannot be attributed to pitch information, since the model orders at which it occurs are below those which might be expected to cause frequency ripple: 30 coefficients derived from a waveform sampled at 10 kHz represent a "quefrency" of 3 milliseconds, while the shortest expected pitch period for a male speaker might be some 5 milliseconds in length.

5.2.3.3. Model orders for poles and zeros

The different response of the pole and zero models to changes in model order revealed in this experiment may be explained in terms of the number of poles and zeros to be expected in the vocal tract transfer function. It is generally considered that the vocal tract transfer function has far *fewer* zeroes than

poles. In many of the studies covered in Chapter 3, for example, the number of spectral peaks exceeds the number of dips: Fant's calculations (1970: 146-7) give four or five peaks in the first 3 kHz for [m] and [n], but only one spectral dip; Nord's (1976b) figures suggest that there are 6 poles in the first 3 kHz for [m] and [n], but only 3 zeros; while Castelli and Badin (1988) found 9 poles and 3 zeroes in the nasal-pharyngeal transfer function.

Zeroes are only introduced by the addition of a shunting cavity – a side-chamber to the main acoustic flow or a chamber behind the sound-source, and each such cavity introduces additional poles at the same time. However, as Atal and Schroeder point out (1978: 1315), zeroes are also introduced into the speech spectrum by the glottal excitation and lip radiation characteristics, and by the use of a low-pass anti-aliasing filter in digitisation.

Thus overestimation of the model order will occur earlier in the all-zero model than in the all-pole model. Overestimation was not thought to be a problem in Chapter Four, since the technique dealt very well with signals produced by low-order digital filters, even when the model order was much larger. The situation appears to be different with real speech, however. A possible explanation is that the surplus coefficients – those which are not needed for modelling the major spectral features of the vocal tract transfer function – begin to model spectral noise, which, as was observed above, can be enhanced somewhat in cepstral analysis because of the logarithmic nature of the operations.

This might not be a problem if our only interest was in the combined model response: Atal and Schroeder (1978), for example, achieved a satisfac-

tory pole-zero representation of nasals using identical numbers of pole and zero coefficients with their technique based on Linear Prediction. However, our interest lies in extracting formant information from the separate responses, and the increased noise makes accurate formant location impossible.

5.2.4. Experiment 2: reducing the order of the all-zero model

It appears, then, that overestimation of the model order for real speech, as opposed to a noise-free synthetic signal, is not without problems if we intend to use the all-pole and all-zero responses separately. The problem was worse for the all-zero signal, since the fact that there will be fewer spectral zeroes to model means that overestimation occurs much earlier. It was therefore decided to alter the pole-zero algorithm to allow a lower order all-zero model to be derived, while deriving the all-pole model with the same number of coefficients as the smoothed negative derivative of phase spectrum. Since the effective separation of the pole and zero responses depends upon using a relatively large number of cepstral coefficients, the smoothed NDPS was derived from the full number of cepstral coefficients (as in the existing algorithm); the positive and negative halves were returned to the cepstral domain by the Inverse DFT; and the resulting all-pole and all-zero cepstral signals were modelled separately using Linear Prediction, deriving the M pole coefficients as in Equation 4.20 (Chapter 4), but limiting the number of zero coefficients j in Equation 4.21 to Z, where $Z<M$. This amounts to smoothing the all-zero cepstral response more than that of the all-pole model. Since the number of cepstral coefficients used to generate the NDPS and the number of all-pole predictor coefficients are

unchanged in each case, the frequency response of the all-pole model – and its degree of separation from the all-zero model – remains the same: only the all-zero model itself, and thereby the combined pole-zero model response, are affected.

5.2.4.1. Acoustic analysis

The modified pole-zero analysis was applied to only the bilabial nasal token used in the first experiment, under the same conditions. The overall model order chosen was 25; the number of pole coefficients derived was 25 in each case, but the number of zero coefficients derived was decreased in steps, from 18 to 12 to 6 to 2.

5.2.4.2. Results

Figure 5.4 shows the effect of the reduced zero model order on the output spectra for the bilabial nasal, using an all-pole model with 25 predictor coefficients (derived from 25 cepstral coefficients), and all-zero model orders of 2, 6, 12 or 18 coefficients. A zero model order of 2 is clearly inadequate, but a model order of 12, just less than half that of the all-pole model, gives an output spectrum containing all the essential features seen in the spectrum of higher order models, without the distracting detail. Atal and Schroeder (1978: 1316) made a similar finding, also in the case of the bilabial nasal, noting that a Linear Prediction model with just 6 zeros and 12 poles gave a result as good as that obtained with 12 zeros and 12 poles.

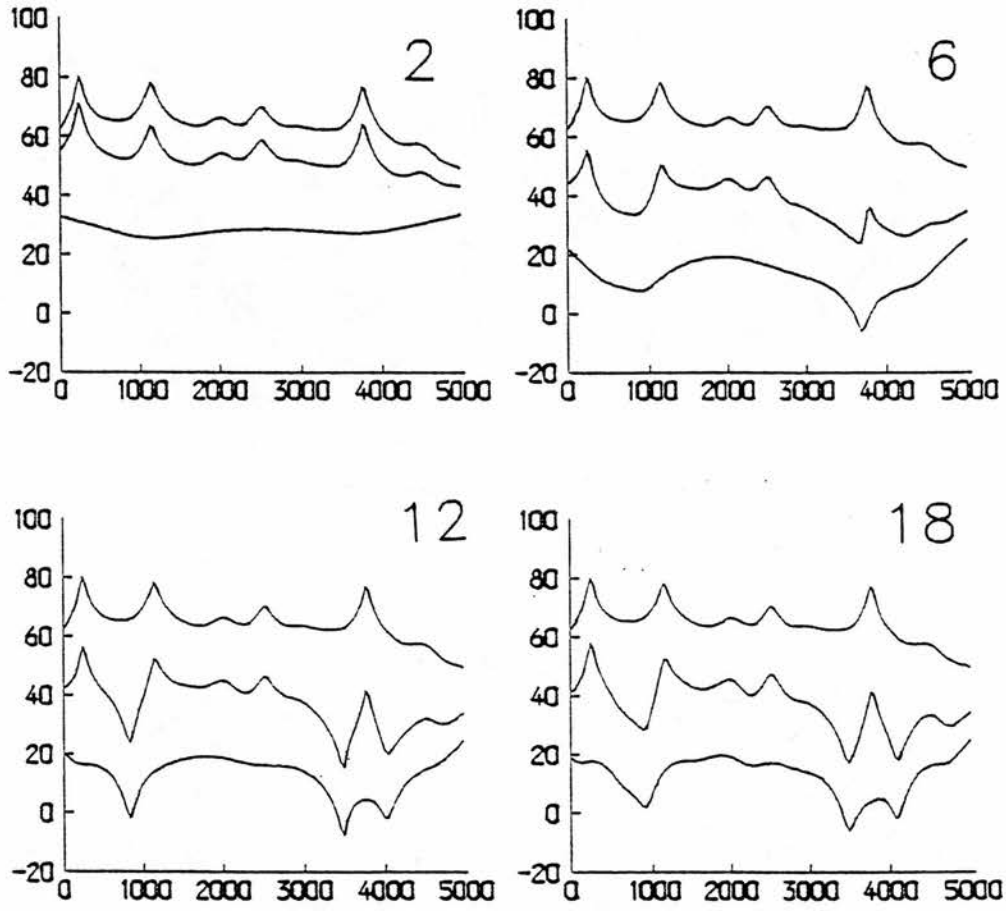

Figure 5.4 Pole-zero spectra for [m] with a fixed number of poles and an increasing number of zeros. Traces: all-pole (top), pole-zero (middle), all-zero (bottom)

5.2.5. Summary of Experiments 1 and 2

The first experiment revealed that overestimation of the model order could be a problem in the case of the all-zero model. A solution was proposed whereby the all-zero model order was lowered, and this was tested in the second experiment. The approach still achieved a satisfactory separation of the all-pole and all-zero responses, and mitigated the noise problem created by

overestimation. At the model orders used here, a zero model order of only half that of the all-pole model proved adequate.

5.3. Formant and anti-formant frequency data for the three nasals

5.3.1. Outline

The aim of the pole-zero analysis so far has been to discover a suitable range of settings for application to speech. This section turns now to the problems of extracting the required acoustic data from the analysis results, in preparation for the analyses which follow in the next chapter. Frequency response spectra for the bilabial, alveolar and velar nasal tokens analysed above are examined, and a peak-picking routine is then applied to the averaged output spectra (all-pole response, all-zero response and combined pole-zero response) to estimate the frequencies and bandwidths of the formants and anti-formants detected. The data obtained are compared with findings presented elsewhere, partly to validate the analysis, and partly to see whether the use of separate pole and zero responses provides any additional information about the spectral structure of nasals.

5.3.2. Acoustic analysis and results

The analysis used the modified version of the pole-zero technique introduced in the last section. The three nasal tokens used in Experiment 1 (sampling frequency 10 kHz) were re-analyzed to generate all-pole models of order 25 and all-zero models of order 12 for each token; all other analysis conditions

remained unchanged. A simple peak-picking routine was then applied to each of the averaged spectra to give estimates of the resonance and anti-resonance frequencies and their bandwidths for comparison with published data. Resonance frequencies were estimated from the *peaks* of the *all-pole* response; anti-resonance frequencies were estimated from the *dips* of the *all-zero* response. Their bandwidths represent the width of the peak (or dip) at a level 3 dB down from the peak on each side.

The averaged all-pole, all-zero and pole-zero frequency response spectra are shown in Figures 5.5, 5.6 and 5.7 ([m], [n] and [ng] respectively).

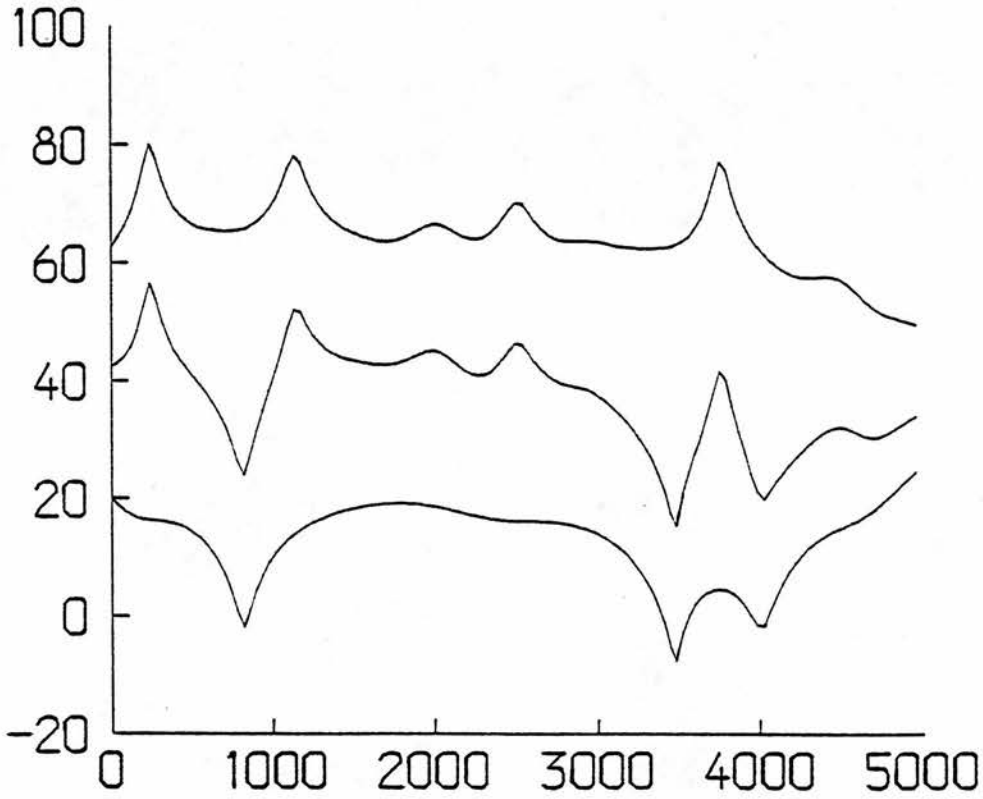

Figure 5.5 Pole-zero spectra for [m], 25 poles, 12 zeros. Traces: all-pole (top), pole-zero (middle), all-zero (bottom)

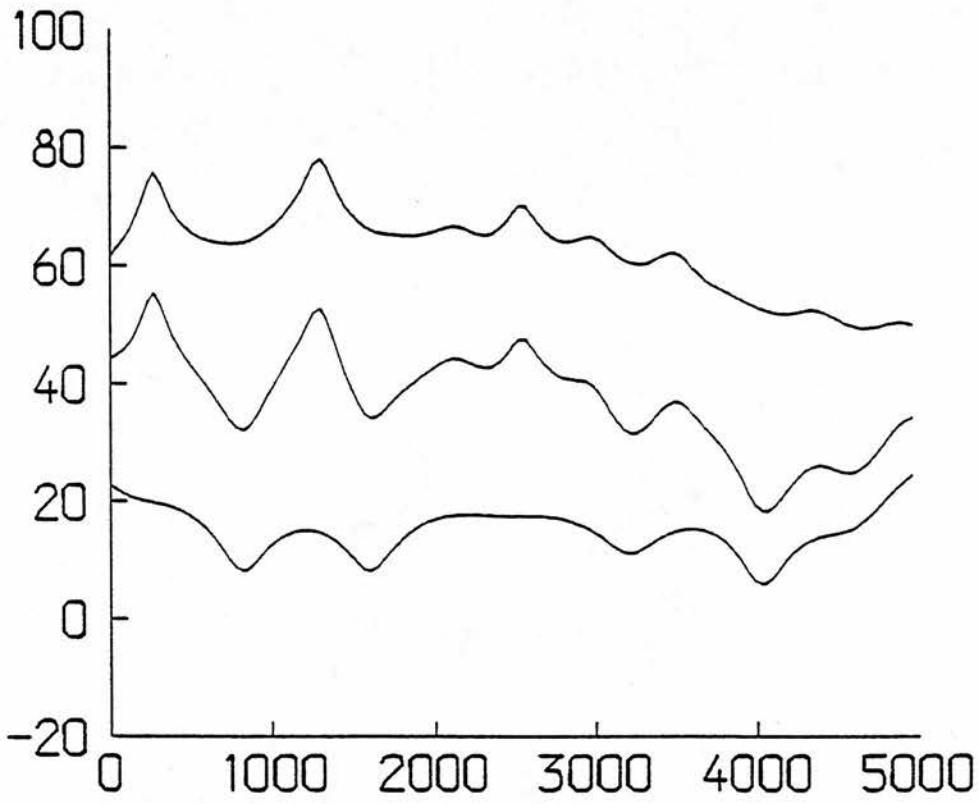

Figure 5.6 Pole-zero spectra for [n], 25 poles, 12 zeros Traces: all-pole (top), pole-zero (middle), all-zero (bottom)

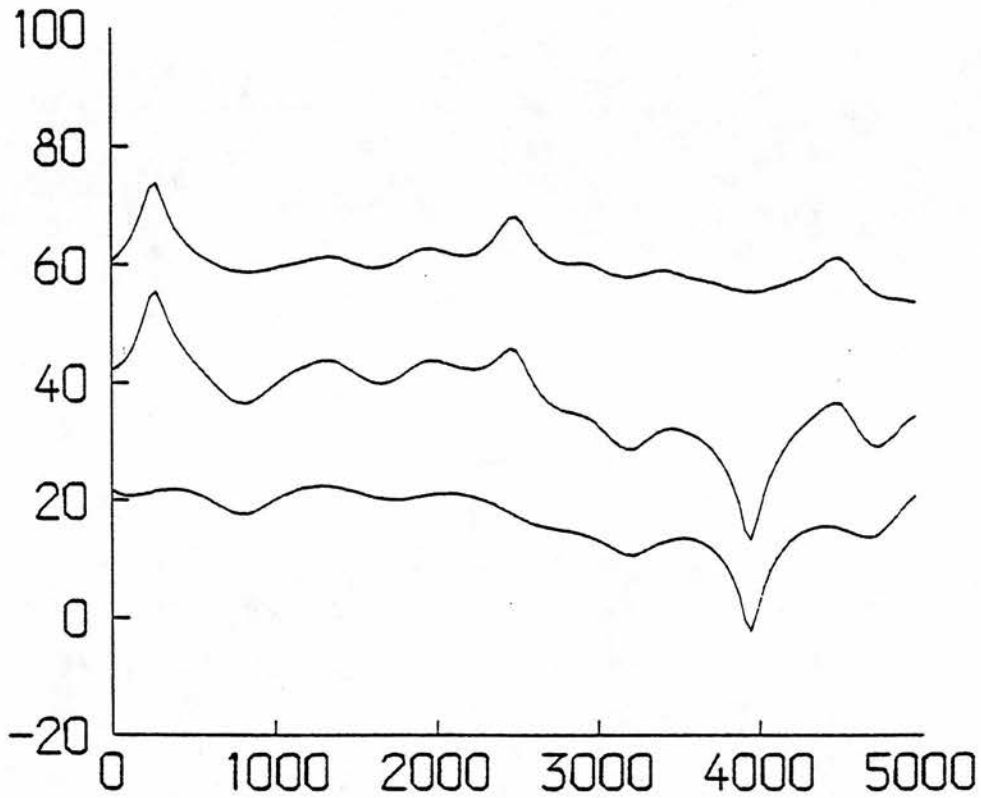

Figure 5.7 Pole-zero spectra for [ng], 25 poles, 12 zeros Traces: all-pole (top), pole-zero (middle), all-zero (bottom)

Table 5.1 gives the frequencies (and bandwidth in brackets) of each peak (dip) for the three tokens.

[m]	peaks	238	1140	2005	2513	2915	3760	4399	
		(88)	(123)	(380)	(193)	(748)	(99)	(585)	
	dips	821	3469	4005					
		(108)	(100)	(159)					
[n]	peaks	273	1287	2112	2557	2953	3457	4327	4881
		(119)	(148)	(458)	(208)	(328)	(352)	(487)	(490)
	dips	825	1591	2478	3198	4029			
		(241)	(252)	(1797)	(338)	(245)			
[ng]	peaks	262	1331	1943	2487	2882	3386	4476	
		(120)	(522)	(494)	(190)	(664)	(523)	(268)	
	dips	107	800	1722	3191	3939	4667		
		(537)	(434)	(758)	(328)	(111)	(429)		

Table 5.1 Peak and dip frequencies and bandwidths for three nasal tokens

5.3.3. Discussion

The *bilabial* nasal (Fig. 5.5) shows the clearest spectral structure, with five prominent formant peaks in the all-pole spectrum (in the first 5 kHz), and three well-defined dips in the all-zero spectrum. The peak detected at 2915 Hz should probably be regarded as spurious, and can be eliminated on the basis of its large bandwidth (748 Hz). The peak at 4399 Hz also has a large bandwidth, but could be related to a peak at a similar frequency seen in the alveolar and velar nasals.

The location of the low frequency anti-resonance is as expected from data in the literature (between 750 and 1250 Hz according to Fujimura 1962, and around 800 Hz according to Fant 1970); a higher frequency anti-resonance (as observed here) was also reported by Fant, at 3500 Hz.

This token also demonstrates the value of estimating resonance and anti-resonance locations from separate all-pole and all-zero responses: the dip visible at 2271 Hz in the combined pole-zero response is clearly the result of the closeness of the two peaks in the all-pole response (at 2005 and 2513 Hz), rather than being due to a spectral zero, since the all-zero response is practically flat in this region.

The *alveolar* nasal (Fig. 5.6) has more resonances, which tend to be broader and, at higher frequencies, much less energetic. The first peak is higher in frequency than that for [m] (as found in data from Recasens 1983), as might be expected from the lower total volume in the vocal tract behind the oral closure (Fant 1970). Its anti-resonances are more numerous and broader in bandwidth than those of the bilabial nasal. The anti-resonance seen at 825 Hz is not mentioned by other authors, who note only the higher frequency features (1800 Hz and 5600 Hz for Fant 1970, 1600 Hz for Fujimura 1962).

The structure of the *velar* nasal (Fig. 5.4.3) is not so clear: the peaks between 1000 and 2000 Hz – well marked for both [m] and [n] – are less prominent and wider in bandwidth than those at 262 and 2487 Hz. This agrees with observations such as that of Hattori et al. (1958), who noted "a dull and complex frequency characteristic" for [ng]. The major anti-resonance for [ng] is

seen at 3939 Hz – hence perhaps the statements of Fujimura (1962) and Fant (1970) that there are no anti-resonances below 3000 Hz in the velar nasal. The features seen below 3000 Hz in this token may be due to residual interaction between the pole and zero models, though in the case of the 800 Hz dip this seems unlikely, since there is no corresponding dip in the all-pole spectrum. The location of this dip indicates that it could be related to the low frequency dip seen in the bilabial (and alveolar) nasals. In the case of the bilabial nasal, this dip is presumably attributable to the oral cavity anti-resonance (Fant 1970: 147); and the same could be true for the velar nasal, whose larger front oral cavity (with the tongue retracted for the velar closure) might explain the slight lowering of the anti-resonance frequency. The origin of this feature in the alveolar nasal remains unclear.

One other notable feature of the velar all-zero response is the very low frequency dip at 107 Hz. Its origin is uncertain, but its large bandwidth suggests that it is an artefact (possibly related to the effects of the fundamental frequency on the FFT spectrum).

5.4. Conclusion: general observations on the pole-zero technique

The pole-zero modelling technique offers a useful tool for estimating both resonance and anti-resonance frequencies of nasal stops. The spectra obtained are a good approximation to the FFT spectrum, and agree broadly with other descriptions of nasal stops. The ability to separate the all-pole and all-zero responses has proved valuable in reducing the number of spurious peaks and dips in the spectrum, since not all peaks (or dips) in the pole-zero spectrum are

attributable to pole (or zero) activity, but can be seen to be caused by activity in the complementary half of the model.

Using a lower modelling order for the derivation of the all-zero model is probably advisable. It allows the interaction between the two halves to be minimized (by the use of a high number of cepstral coefficients in the initial NDPS derivation), while reducing the effects of model overestimation on the all-zero frequency response.

A possible weakness of this study is that the spectral features are derived from the *power spectra* of the separate frequency responses, rather than being obtained directly from the predictor coefficients by root-solving (see Chapter Four, 4.2.1). It is clear from the analysis presented in 5.3 that not every such feature unequivocally represents a vocal tract resonance or anti-resonance, however, and it is probably desirable to set an upper limit – perhaps 500 Hz (Witten 1984) – on the *bandwidth* of spectral peaks and dips, and maybe also a lower limit on their *frequency* (to exclude fundamental frequency effects).

In the next Chapter, the cepstral decomposition technique is used to assess the extent of between- and within-speaker variability, as a prelude to its use in automatic speaker verification in Chapter Seven.

CHAPTER SIX

A STUDY OF VARIABILITY IN THE SPECTRUM OF THE VELAR NASAL STOP

CHAPTER SIX

A STUDY OF VARIABILITY IN THE SPECTRUM OF THE VELAR NASAL STOP

6.1. Introduction

This chapter investigates the extent of inter-speaker and intra-speaker variability in the spectrum of the velar nasal stop. It was suggested in Chapter Three that the velar nasal offered the best potential for segment-based speaker verification, on the grounds that, of the three nasal phonemes in English it was the most resistant to coarticulation effects and the effects of oral cavity anti-resonances. The use of a method of accurately estimating both resonance and anti-resonance frequencies was investigated in Chapters Four and Five. This method of *cepstral decomposition* is now applied to a database of thirty speakers, in preparation for a series of verification trials using the same database in Chapter Seven.

The tokens of the velar nasal are characterised here using the decomposed all-pole and all-zero frequency response spectra. Each token is represented by two vectors comprising the frequencies of the *maxima* in the averaged all-pole response and the *minima* in the averaged all-zero response, as in the analysis of the nasal stops in Chapter Five. Since the number of such maxima may

vary from speaker to speaker, and from token to token, a method is proposed for obtaining proper alignment of corresponding peaks and dips across tokens. The effects of several factors which influence the velar nasal spectrum and affect the performance and reliability of speaker verification systems are then considered: phonetic context (specifically, the identity of a preceding vowel), speaker sex and, crucially for speaker verification, intra-speaker and inter-speaker variation over a period of weeks.

6.2. Materials

The materials for the experiments in this chapter were obtained from recordings of isolated words containing the velar nasal /ng/ in word-final position. The words were recorded by thirty speakers (15 males, 15 females) in eight sessions spread over several months. The velar nasal from each token was isolated by hand for acoustic analysis.

6.2.1. Word lists

The list of words used in the recordings is presented in Appendix B. All words began with a non-nasal consonant or consonant cluster and ended in the velar nasal /ng/, with one of the three vowels /i/, /a/, /uh/ (or /u/ in the case of Nthn.Eng. speakers) immediately preceding the nasal; these particular vowels were chosen as being the most peripheral of the restricted set of vowels – /i, a, uh (u), o, e/ – which can occur in British English before /ng/ in the same syllable (Gimson 1970). The initial consonant or consonant cluster was varied to avoid having too many repetitions, and to exclude the use of nonsense words,

but was not otherwise controlled.

For reasons associated with the preparation of database materials, the words used in the first session differed from those used in the remaining seven sessions and fewer tokens were provided. Their structure was similar, however, and the differences are assumed to have no effects on accuracy.

6.2.2. Speakers

Speakers were colleagues from the Department of Linguistics and the Centre for Speech Technology Research at Edinburgh University. They were chosen for their availability to make recordings at various times of the day over several weeks. A variety of accents were represented, including Standard Scots, Lothians, Northern English, English Received Pronunciation and (in one case) General American English. Speakers' ages ranged from 19 to 39 years (mean 27.7 years) for males, and from 18 to 55 years (mean 31.6 years) for females.

The composition of this group, whilst not exhibiting true homogeneity for age and accent, does represent a valid cross-section of the type of speech community which might be required to use a speaker verification system.

6.2.3. Recording

Each speaker made eight sets of recordings (with one exception: speaker ED374F provided only four), spread over a period of at least eight weeks in each case. Recordings were made at varying times of day, and no more than two sessions were recorded on any one day by a given speaker.

Recordings were made in a sound-treated room using high-quality digital equipment (Sony PCM-F1 digitiser (14-bit), Sony SL-F1E digital recorder and Sennheiser MKH-406 directional condenser microphone). The speaker sat at a table facing a colour TV monitor. Instructions and reading materials appeared on the screen of the monitor under the control of the operator, by means of a program for storing and displaying prompts and text on a microcomputer. A variety of materials including short sentences, lists of isolated words and a long reading passage, was provided for the general database recording, but only the isolated words containing the velar nasal were used in these experiments.

6.3. Analysis of nasal tokens

Each nasal token was processed using the cepstral decomposition method as described in Chapter Four and modified in Chapter Five. The speech was first downsampled from 20 kHz to 10 kHz, to allow analysis at a lower order to be used and to eliminate any possibility of spectral aliasing in the higher frequencies. Investigation of selected spectra confirmed that there were no significant spectral features above 5 kHz.

The same analysis parameters were used for both male and female speakers. A shorter analysis window is sometimes recommended for female speakers, since their higher fundamental frequency gives shorter pitch periods, and the requirement for at least two pitch periods can be met with a much shorter speech interval. However, use of a shorter length Fourier transform degrades the frequency resolution of the analysis. Use of a long analysis interval brings with it the problem that the vocal tract position may change during the

analysis interval, causing the details of the output spectrum to be distorted. However, there is no evidence that female speakers show a higher rate of change of vocal tract configuration than males, and therefore no proven reason not to use the same window length. To obtain the same frequency resolution in the output spectra, and to allow the same model order for all speakers, therefore, it was decided to use exactly the same window length and transform size for each sex. For the work here, each window, or frame, of speech was of 25.6 ms duration (256 sample points).

Each frame of speech was multiplied by a Hanning window to reduce discontinuity effects between frames. Frames were then submitted to the pole-zero analysis. Spectral pre-emphasis of 0.97 was applied, and a 256-point FFT used to obtain the cepstrum coefficients. The first 25 cepstral coefficients were used to generate a 25th order all-pole model and a 12th order all-zero model, as described in Chapter Five. The frequency response of each half of the model – the *all-pole* and *all-zero* spectra – was then derived. Each spectrum covered the frequency range 0 to 5 kHz in 128 points, giving a frequency resolution (the interval between points) of 39 Hz.

This analysis was repeated for each frame of data in a token, with a frame shift of 5 ms, so that frames overlapped. The output spectra were then averaged over the duration of the token to give the average all-pole, all-zero and pole-zero spectra; averaging was done on the log power spectra because a smoother result was obtained than if the linear power spectra were used.

6.3.1. Location of spectral features

Each nasal token was represented by a small set of spectral features corresponding to formant and anti-formant frequencies. These features comprised the maxima of the averaged all-pole spectrum and the minima of the averaged all-zero spectrum. They were obtained using a simple peak-picking algorithm. Peaks below 160 Hz for males and 200 Hz for females were excluded, to eliminate any features relating to fundamental frequency which occasionally appeared in the output spectra despite the low model order used (cf. Chapter Five, 5.3.3). Dips below 200 Hz were excluded to avoid the possibility of detecting features related to the spectral pre-emphasis filter applied during the analysis. In addition, any peaks or dips with a bandwidth greater than 500 Hz were removed: this value is suggested by Witten (1984: 84) as a suitable limit on the bandwidth of genuine speech formants, and has been used in formant tracking algorithms (e.g. Markel and Gray 1976).

6.4. Preliminary analysis

6.4.1. Numbers of peaks and dips found

Figure 6.1 shows the distribution of the numbers of peaks detected in the all-pole spectra of all fifteen male and fifteen female speakers (within the frequency and bandwidth limits described above). Figure 6.2 shows the distribution of the numbers of spectral dips in the all-zero spectra of the same speakers. There is a measured tendency for female speakers to have lower numbers of both peaks and dips than male speakers: the female speakers show a lower

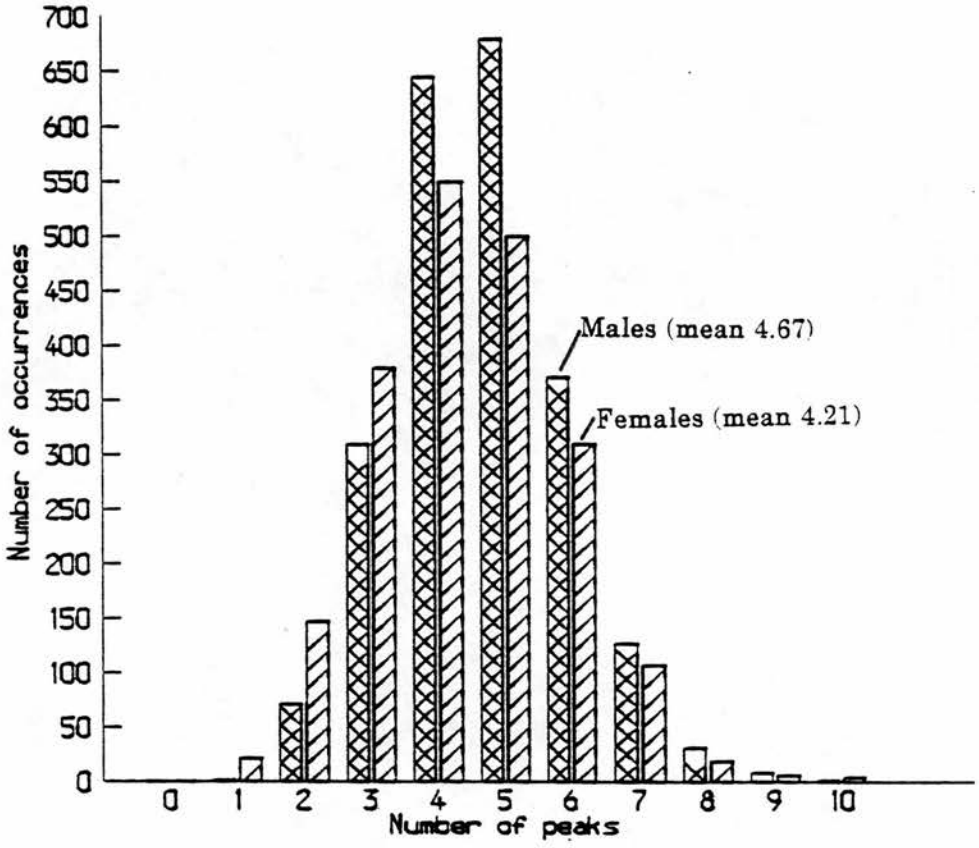

Figure 6.1 Numbers of spectral peaks per token for 15 male and 15 female speakers

mean and mode in each case, and chi-squared tests confirm that the distributions are significantly different (chi-squared values of 86.4 for the peaks and 103.1 for the dips, both significant at $p < 0.001$).

It is to be expected that females should show somewhat fewer spectral peaks in the first 5 kHz, since much of the spectral structure of velar nasals can be attributed to the resonance characteristics of the nasal-pharyngeal tube (Chapter Three), which depend partly on its length: a shorter nasal-pharyngeal tube produces resonances which are higher in frequency and more widely

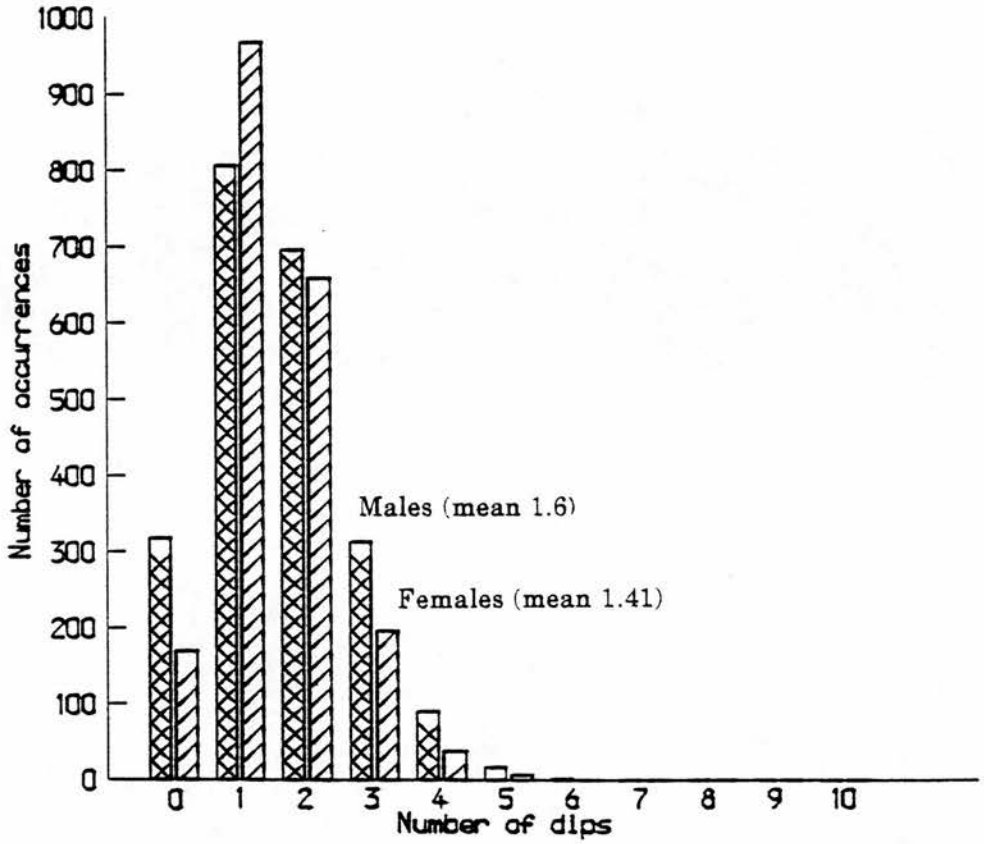

Figure 6.2 Numbers of spectral dips per token for 15 male and 15 female speakers

spaced, and therefore fewer in number within a fixed frequency range. However, the different numbers of peaks could also indicate greater difficulty in locating peaks in the spectrum for female speakers (owing to higher bandwidths, for example). Similar difficulties might explain the slight reduction in the numbers of dips in the spectrum for the female speakers.

The numbers of peaks and dips found in tokens according to their vowel context are presented in Figures 6.3 to 6.6.

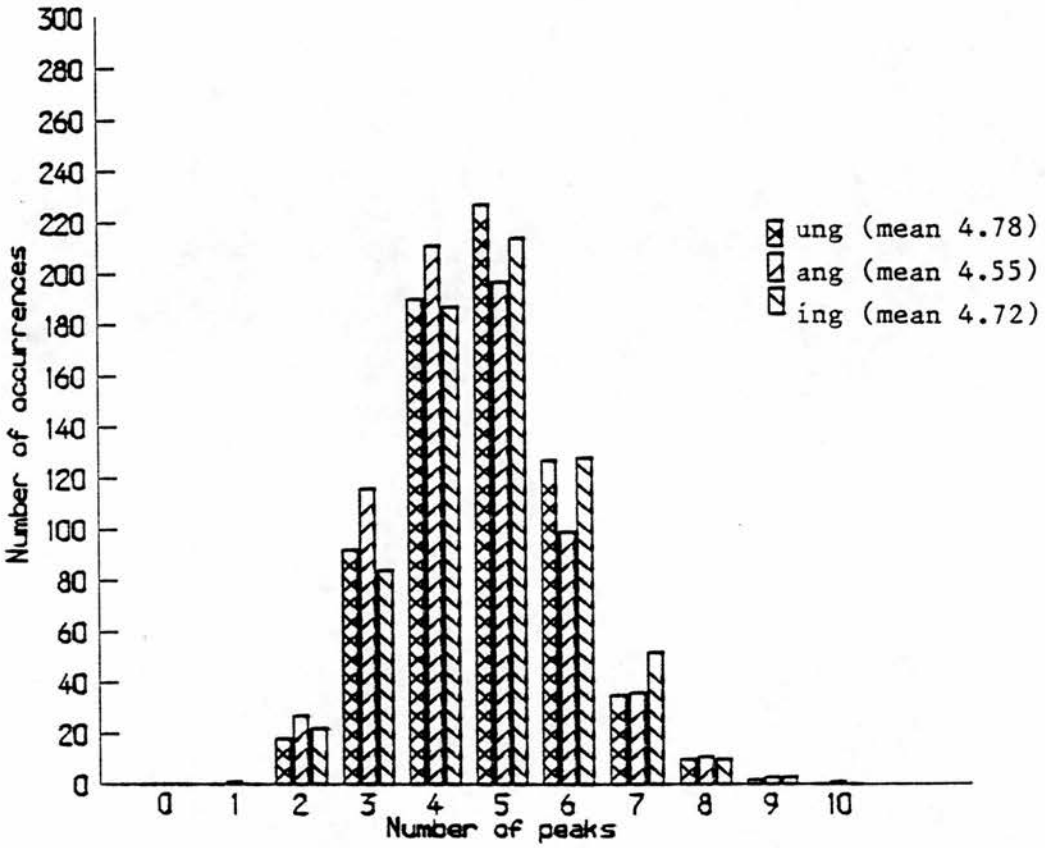

Figure 6.3 Numbers of spectral peaks per token by vowel context (males)

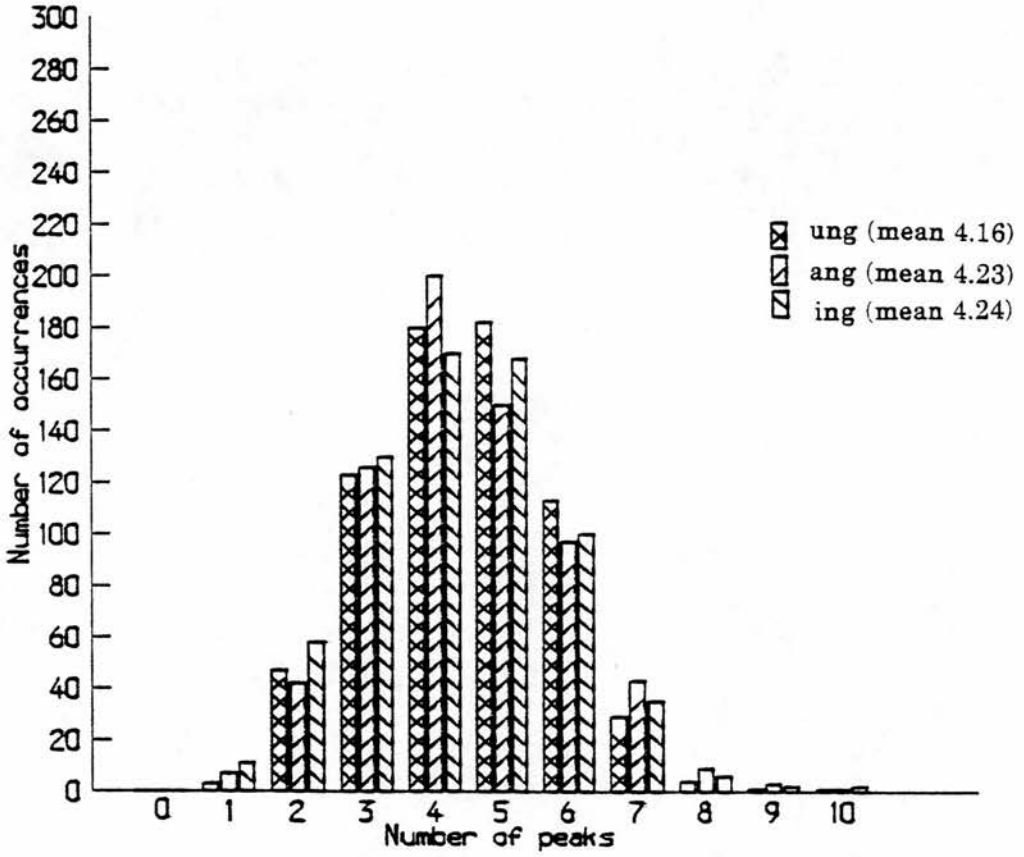

Figure 6.4 Numbers of spectral peaks per token by vowel context (females)

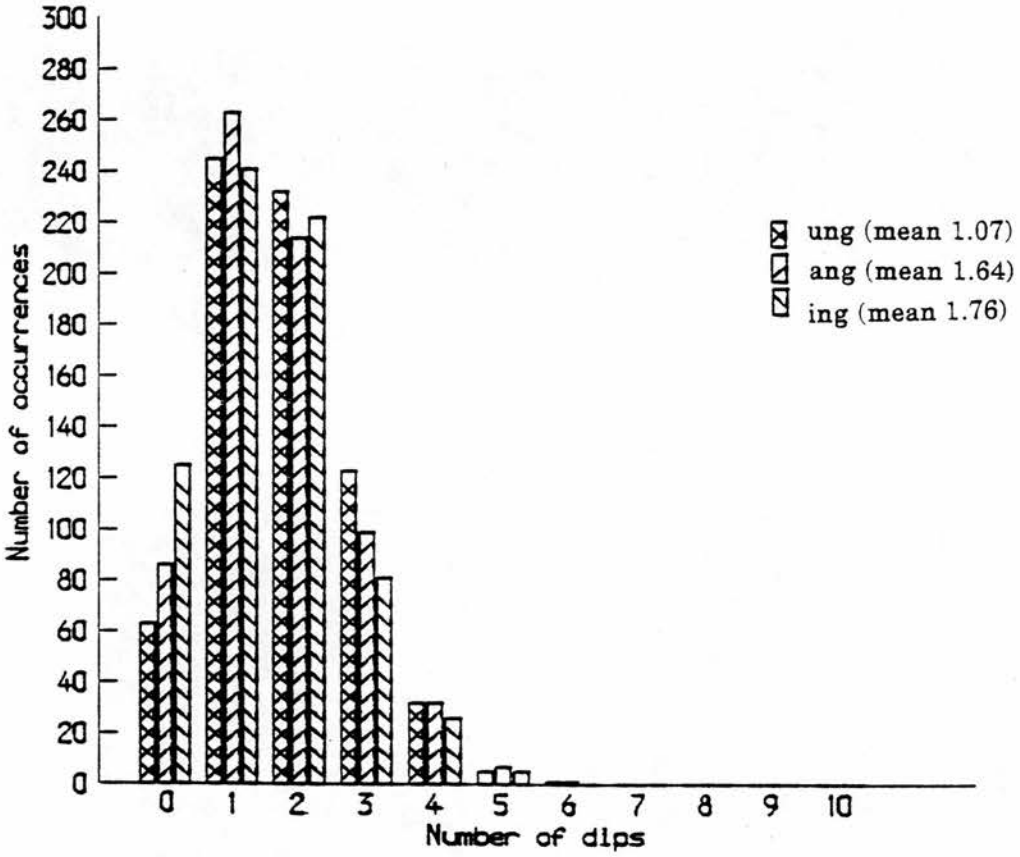

Figure 6.5 Numbers of spectral dips per token by vowel context (males)

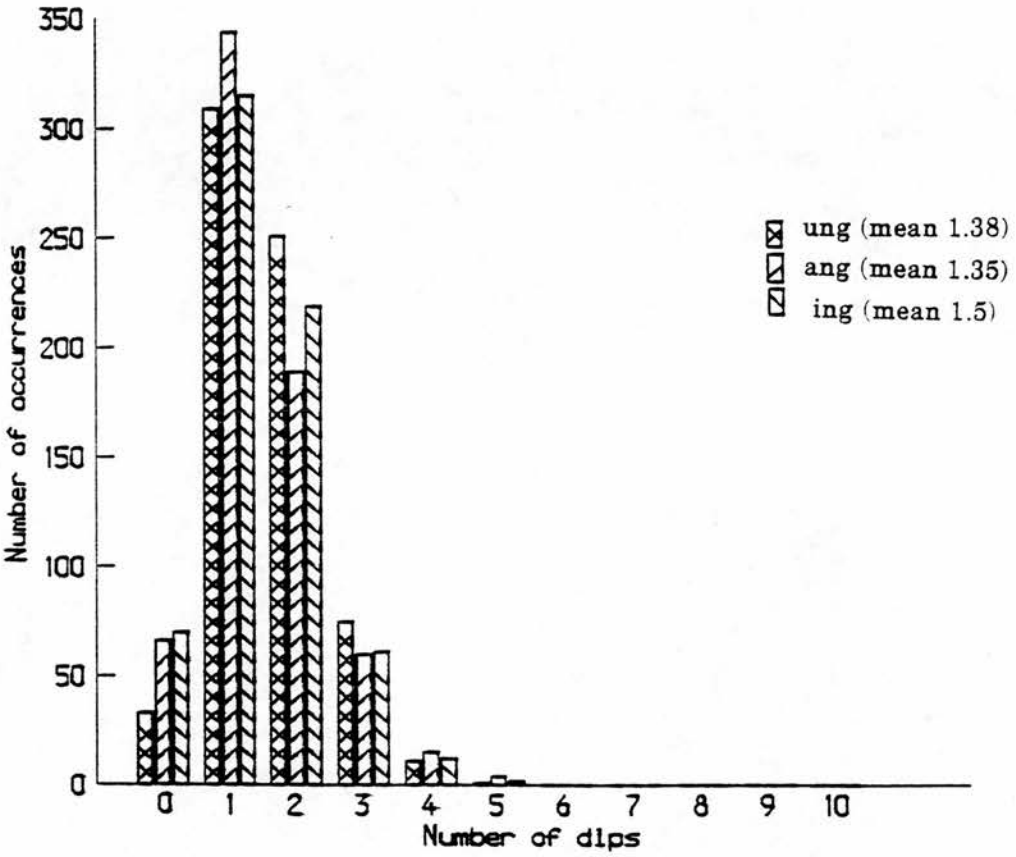

Figure 6.6 Numbers of spectral dips per token by vowel context (females)

Each bar in these charts represents the number of tokens yielding that number of peaks (or dips) in one of the three vowel contexts (after /i/, after /a/ and after /uh/). There appear to be no consistent differences among the vowel contexts in the numbers of peaks and dips detected.

6.4.2. Peak and dip frequencies

It was seen in the preceding section that the *numbers* of peaks and dips detected in the spectrum varied significantly for both male and female speakers. In this section, the distribution of peaks and dips on the *frequency* scale is considered. Data for "Peak 1" (P1) relate to the first detected peak in each token, those for "Peak 2" (P2) the second detected peak, and so on.

Figures 6.7 and 6.8 show the distribution of peak frequencies for the fifteen male speakers and the fifteen female speakers respectively; these distributions are summarized in Table 6.1. Similarly, Figures 6.9 and 6.10, and Table 6.2, present the distributions of the spectral dip frequencies for the thirty speakers.

Parameter		Males	Females
P1	mean	308.6	270.1
	s.d.	(255.8)	(91.7)
	N	1876	1834
P2	mean	1473.4	1343.0
	s.d.	(645.0)	(598.1)
	N	1896	1821
P3	mean	2283.5	2240.7
	s.d.	(742.0)	(811.6)
	N	1849	1696
P4	mean	2999.9	2998.2
	s.d.	(808.3)	(853.8)
	N	1583	1322
P5	mean	3582.3	3577.3
	s.d.	(712.9)	(782.4)
	N	1040	851
P6	mean	3931.1	4019.3
	s.d.	(593.3)	(654.2)
	N	454	398
P7	mean	4062.6	4221.3
	s.d.	(467.6)	(568.0)
	N	147	123
P8	mean	4159.3	4307.1
	s.d.	(393.1)	(331.4)
	N	37	26
P9	mean	4218.5	4504.3
	s.d.	(373.3)	(323.4)
	N	9	8
P10	mean	4536.4	4534.9
	s.d.	(0.0)	(494.0)
	N	1	2

Table 6.1 Distributions of raw spectral peak frequencies (Hz)

Parameter		Males	Females
Z1	mean	1330.7	1036.4
	s.d.	829.7	539.1
	N	1664	1680
Z2	mean	2576.9	2211.2
	s.d.	851.8	775.2
	N	980	810
Z3	mean	3281.4	2879.9
	s.d.	741.2	783.4
	N	355	221
Z4	mean	3777.0	3255.1
	s.d.	576.9	760.5
	N	104	43
Z5	mean	3965.3	3977.7
	s.d.	366.5	488.2
	N	16	7
Z6	mean	4424.2	-
	s.d.	525.1	-
	N	2	-

Table 6.2 Distributions of raw spectral dip frequencies (Hz)

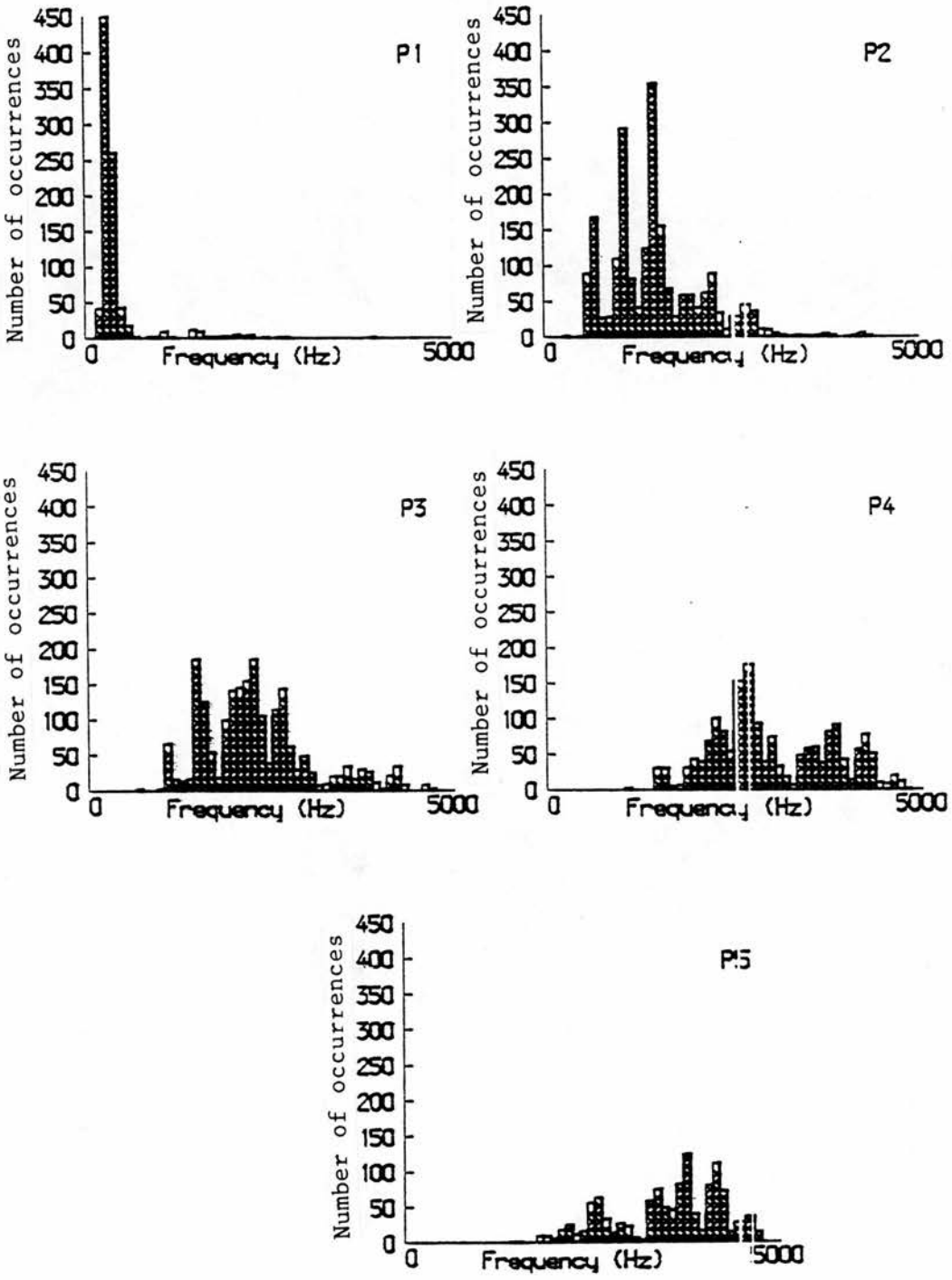

Figure 6.7 Pole frequency distributions (raw data) for 15 male speakers

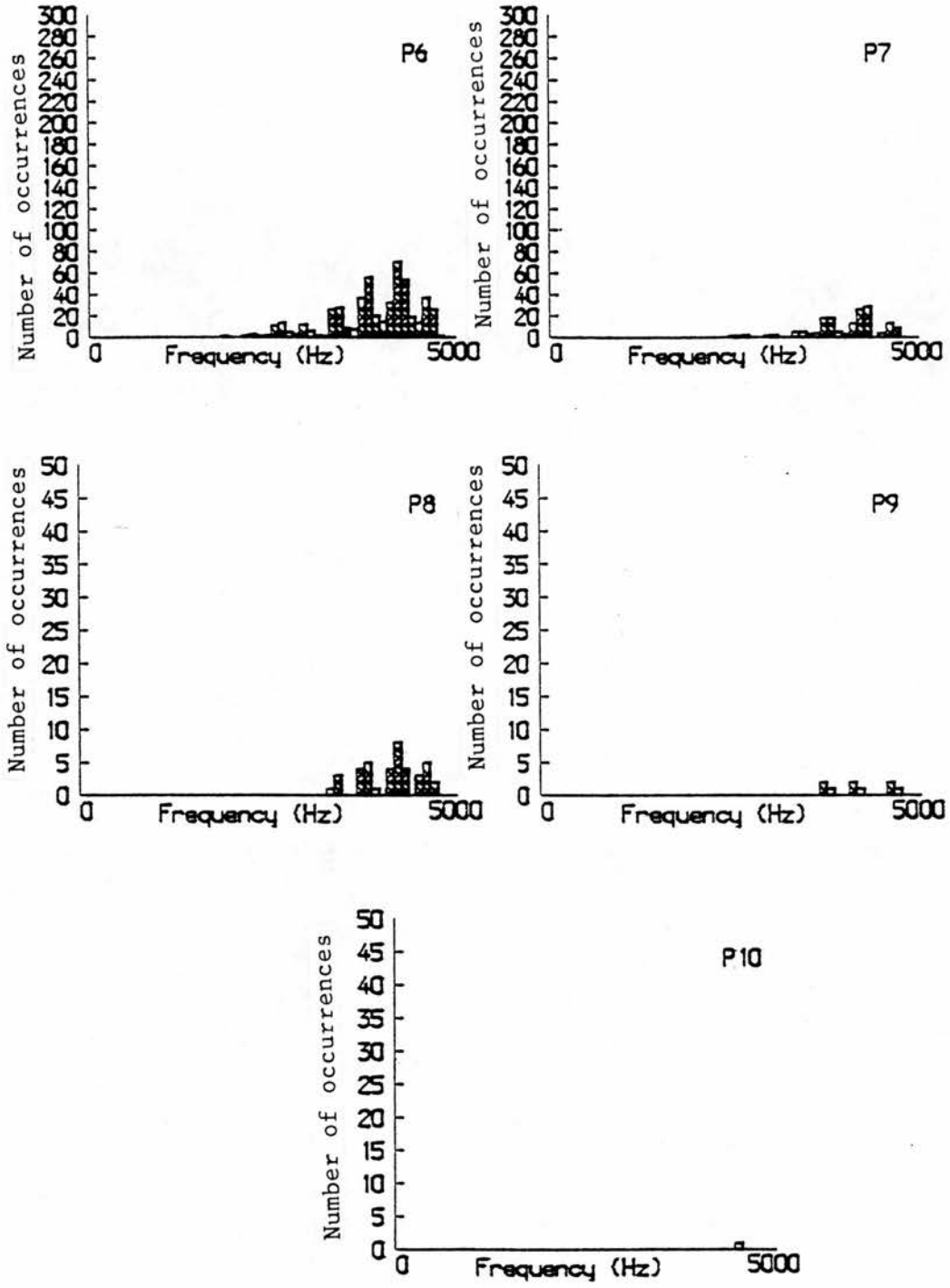

Figure 6.7 Pole frequency distributions (raw data) for 15 male speakers

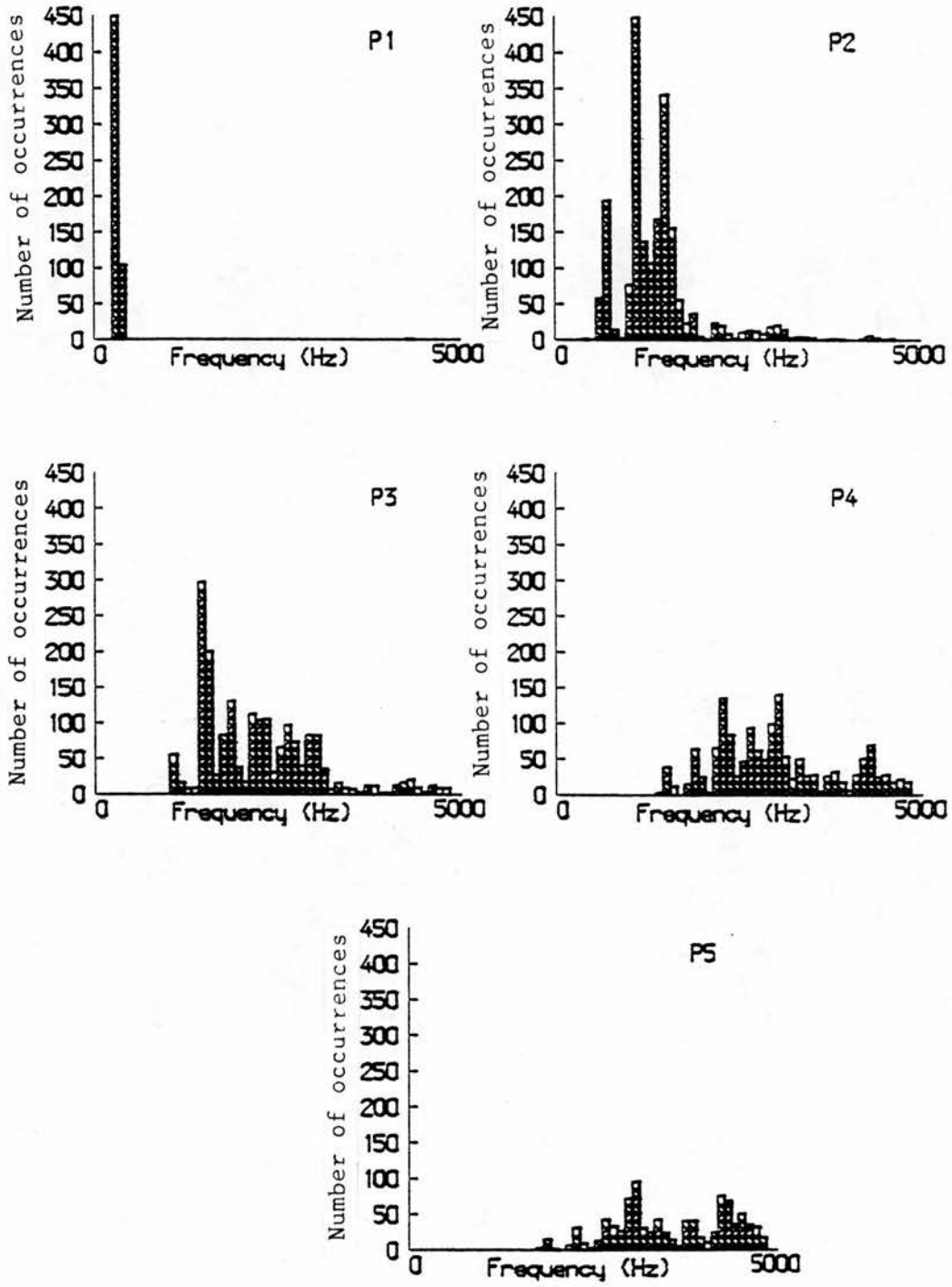

Figure 6.8 Pole frequency distributions (raw data) for 15 female speakers

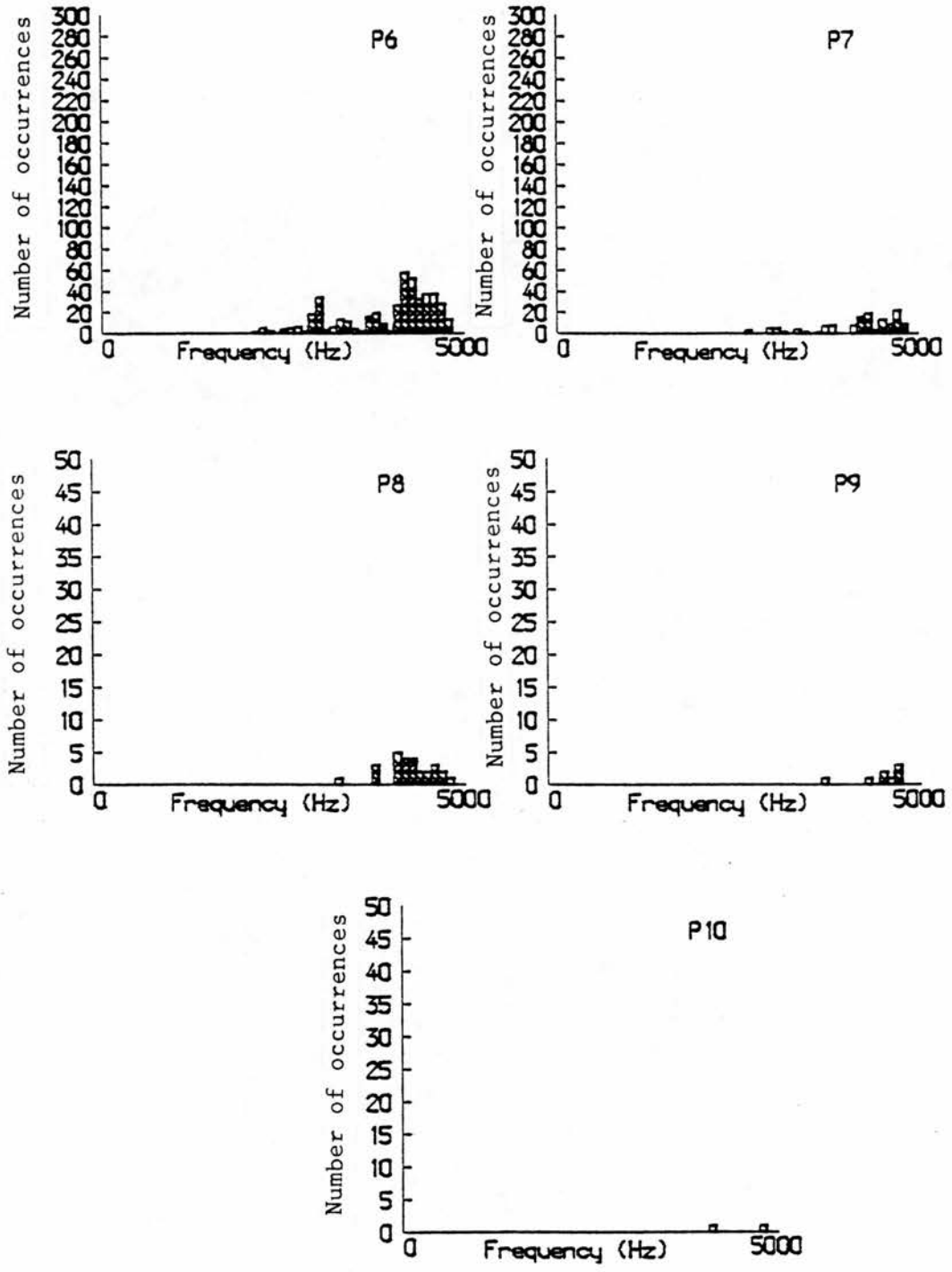

Figure 6.8 Pole frequency distributions (raw data) for 15 female speakers

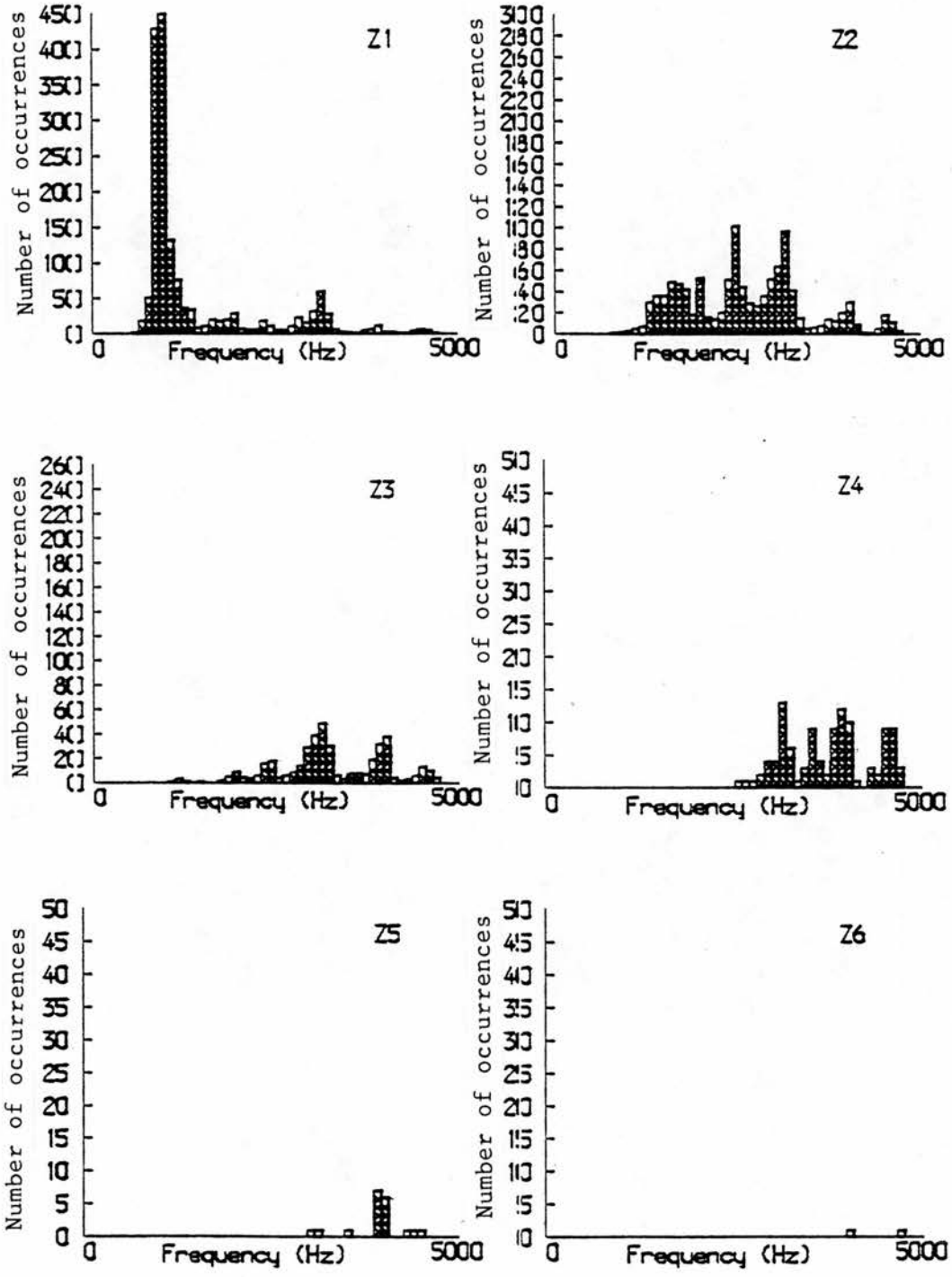

Figure 6.9 Zero frequency distributions (raw data) for 15 male speakers

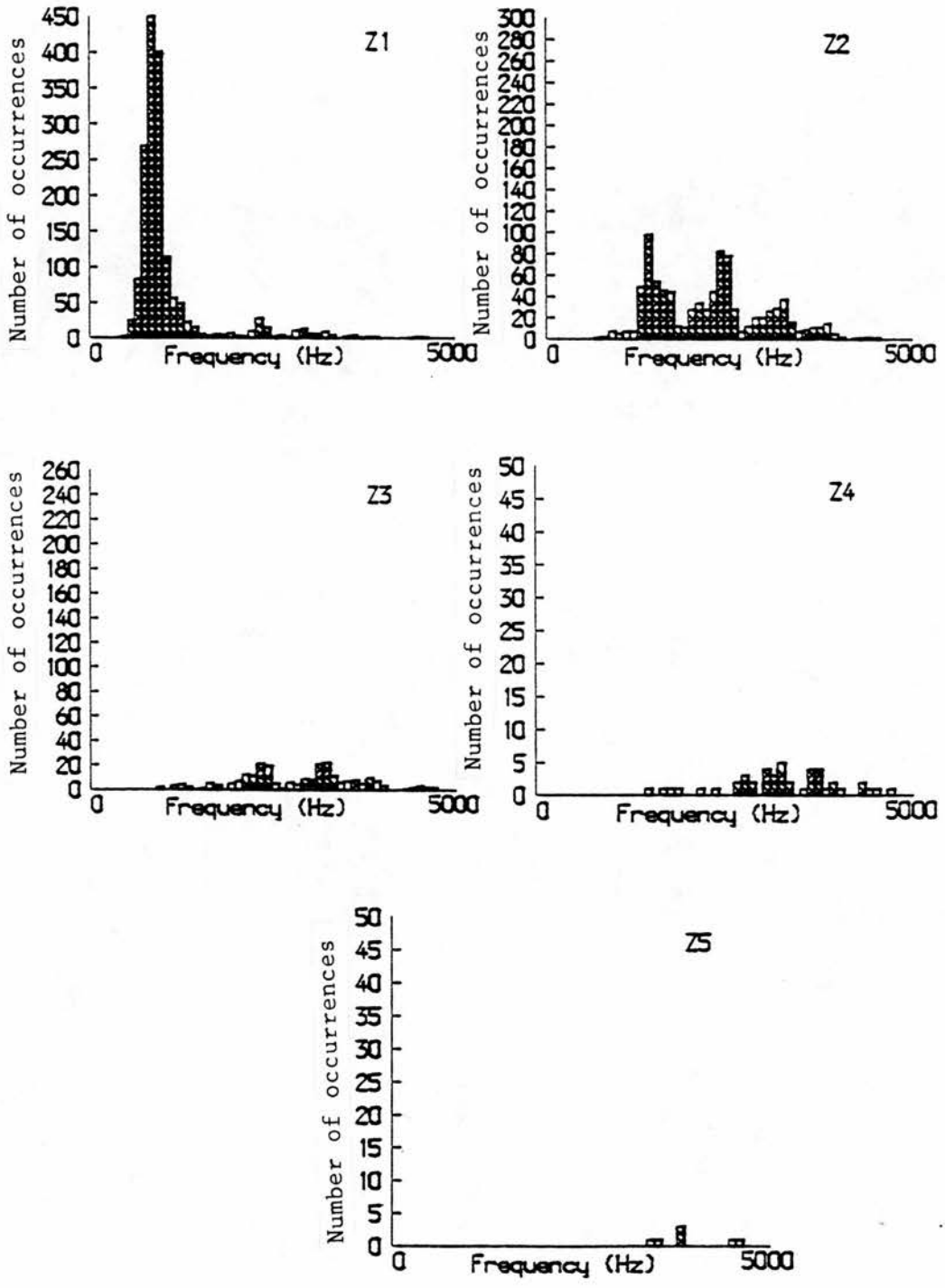

Figure 6.10 Zero frequency distributions (raw data) for 15 female speakers

It is clear that all peak and dip frequencies show very wide distributions, and in several cases these distributions are multi-modal and far from normal. In addition, the ranges of consecutively numbered peaks overlap considerably. Closer examination reveals that the same modal frequencies occur in distributions of several of the peaks, suggesting that peaks may in certain cases be assigned wrongly to individual categories. The example presented in Figure 6.11 and Table 6.3 illustrates this effect.

Figure 6.11 shows the all-pole spectra obtained from the first five tokens of /ng/ for the first male speaker ED364M, all spoken in the vowel context /uh/, with the locations of the peaks marked by ordinal number. Table 6.3 gives their frequencies. In each case the last peak is remarkably consistent in frequency, at around 3800 Hz, but in the first token it is assigned to peak 4, since only four peaks were found in that token. Similarly, peak 2 of the first token, at 2171 Hz, appears to have more in common with peak 4 of the second and peak 3 of the third, fourth and fifth tokens. The omission of certain peaks in some tokens and the inclusion of extra peaks in other tokens is thus distorting the distributions of peaks (and dips), not only across speakers (who might be

Peak no.	A	B	C	D	E
1	262	237	246	223	250
2	2171	1065	1598	1619	611
3	2525	1487	2162	2154	2184
4	3887	2176	2542	2492	3418
5		3880	3844	3837	3832

Table 6.3 Peak frequencies (Hz) of spectra for ED364M

expected to show rather different peak distributions) but also within the performance of a single speaker, in a single vowel context. A similar apparent mismatch of formant peaks can be seen in the data presented by Kurowski and Blumstein (1984), summarised in Chapter Three (Table 3.1).

The problem of the correct identification of observed peaks with the correct formants has been commented on by other researchers. The problem is typically one of excluding *spurious* peaks ("energy peaks which may be quite marked, but which do not fit into the normally found formant pattern for the

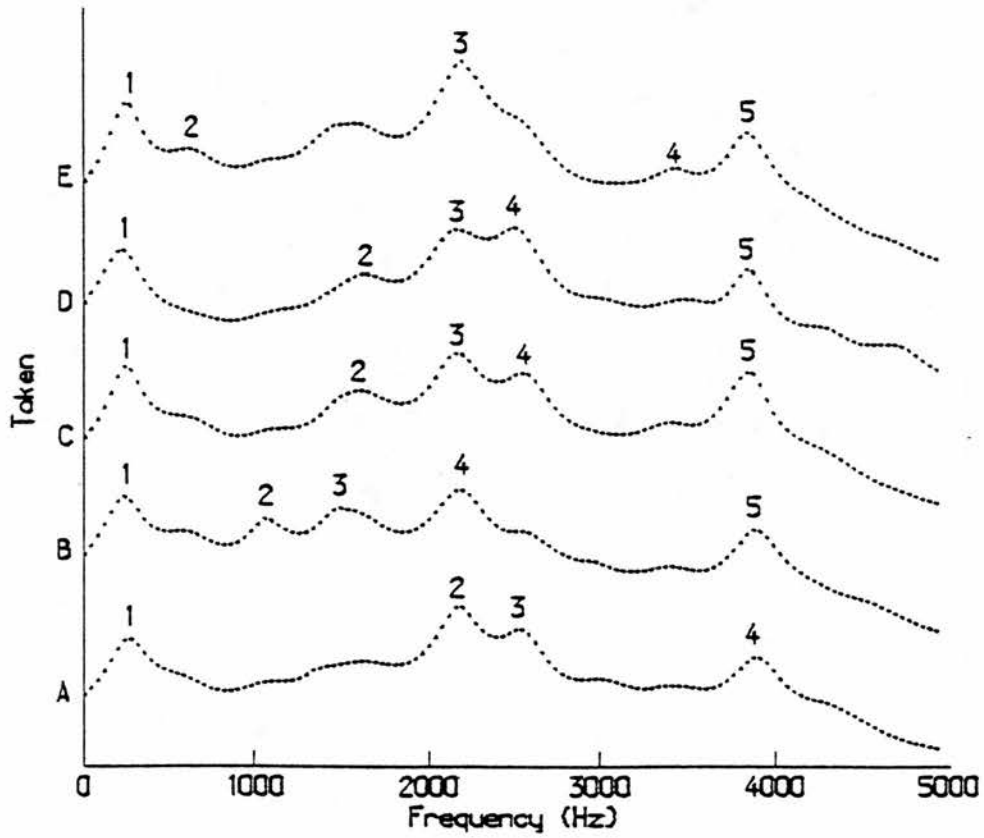

Figure 6.11 All-pole spectra for speaker ED364M showing peak locations

sound" – Nolan 1983: 86). Nolan (1983) suggests that the best remedy for these is to impose on the raw formant data a pattern based on theoretical predictions of where formants should lie, and prior knowledge of the sounds under investigation. The approach adopted by Beddor (1983) was based on judgments of the distinctiveness of formant peaks and their agreement with previous reports in the literature. In both cases, then, labelling of spectral peaks as belonging to individual formants was done subjectively, and involved mainly the exclusion of extra, unwanted peaks.

In the present study, however, such subjective labelling of formant peaks was considered inappropriate because the problem is not only one of excluding spurious peaks, but one of supplying peaks which appear to be *missing*. As Chapter Three has shown, there are very few detailed descriptions of the spectral patterns expected for the velar nasal, other than that there should be approximately four peaks in the first 3 kHz and perhaps one spectral dip just above this frequency. It is also to be expected that speakers will vary in the frequency ranges of individual spectral peaks, and a pattern appropriate for one speaker will probably not be appropriate for others. A more objective method of correctly assigning peaks to individual categories is therefore required.

6.5. A method for obtaining optimal peak alignment

The method proposed here for obtaining optimal alignment of peak and dip profiles has two parts: the establishment of a prototypical vector, defining the pattern of frequencies to be regarded as normal; and the warping of the observed peak vectors to fit this pattern.

Consider again the peak profiles presented in Table 6.3 and Figure 6.11. Let us assume that there is an underlying pattern of resonance peaks for this speaker, and that this pattern can be seen in at least some of the tokens; this is only an assumption, since it is possible that the underlying pattern is never actually manifested in any one token, but with a large enough sample of tokens this becomes less likely to occur. Let us also assume that one of these profiles is a good approximation to this underlying pattern, and can be used as a prototype. It is not clear which this should be, since each contains peaks which do not occur in other profiles. There are consistent patterns, however, which should presumably form part of the prototype: a peak at around 250 Hz, another around 2100 Hz and the last at around 3800 Hz, for example. It seems reasonable that the more frequently a peak occurs, and the more consistent it is in its value, the greater is the probability that it represents an underlying peak for that speaker.

Let us assume for now that profile (D) represents the underlying pattern or prototype. We could then align the other profiles with that prototype by matching the peaks in frequency as in Table 6.4. This still causes some

Peak no.	A	B	C	D	E
1	262	237	246	**223**	250
2		1065/1487	1598	**1619**	611
3	2171	2176	2162	**2154**	2184
4	2525		2542	**2492**	3418
5	3887	3880	3844	**3837**	3832

Table 6.4 Peak frequencies for ED364M aligned with token D

anomalies, however: token A is apparently missing its second peak, token B has one peak too many in second place and no fourth peak, while token E has peaks at positions 2 and 4 which do not match any of the other tokens, let alone the prototype. There seems to be a case, then, for allowing some peaks to be omitted if they do not match the peaks of the prototype, and, conversely, a need for the insertion of "missing" peaks where the prototype has a peak not found in the other profiles. A method for achieving this is proposed here.

6.5.1. Peak profile warping

The approach adopted here is based loosely on the methods used successfully in Automatic Speech Recognition for obtaining the best *temporal* alignment of analysis frames from different speech tokens. A common problem in speech recognition is that tokens of the same word spoken on different occasions may be of different lengths. This causes errors when comparing spectral representations of the two tokens, since analysis frames being compared may come from different parts of each utterance, while frames at the end of the longer utterance will be ignored. The usual solution to this problem is a technique known as *Dynamic Time Warping* (Rabiner et al. 1978, Sakoe and Chiba 1978): each utterance is non-linearly "stretched" or "compressed" by duplicating or omitting analysis frames as necessary to achieve the best match. This is accomplished by calculating a matrix representing the distance between each frame of the reference token and every frame of the test token, and choosing the path through this matrix which gives the lowest total accumulated distance. This distance gives an idea of the quality of the match.

In this thesis a similar method is proposed to decide on the optimal alignment between the prototype peak vector and the "test" vector. The distance measure used is the squared Euclidean distance (see Chapter Seven, Eqn. 7.1), defined as the sum of the squared differences between the elements of each vector. The lower this quantity, the better is the match between the two.

As in DTW, the method begins by calculating a matrix of distances (squared differences) between each element of the prototype vector and every element of the other vector, as illustrated in Table 6.5. A path is then traced through this matrix to find the set of pairs of elements which give the lowest sum of distances. The search for the path begins at the point of closest correspondence between the two vectors: that is, at the matrix cell containing the lowest distance. The search then moves out in each direction from this point to the next cell – diagonally, vertically or horizontally adjacent – with the lowest distance; movements proceed upwards and to the left, and downwards and to the right, until the the two corners of the matrix are reached.

Prototype vector	Test vector				
	250	**611**	**2184**	**3418**	**3832**
223	**729***	**150544**	3845521	10208025	13024881
1619	1874161	1016064	**319225**	3236401	4897369
2159	3644281	2396304	**625***	1585081	2798929
2492	5026564	3538161	**94864**	857476	1795600
3837	12866569	10407076	2732409	**175561**	**25***

Table 6.5 Distance matrix showing optimal warping path between prototype and test vector

The result is a path stretching from the upper left corner of the matrix (pairing the first element in each vector) to the bottom right corner (pairing the last element in each vector). In the event of a tie between adjacent cells of the matrix, diagonal moves are preferred to horizontal or vertical moves, since this results in a lower total distance between the vectors.

Where the path runs horizontally along a row (e.g. row 1, Table 6.5), it indicates that more than one element in the test vector provides a match for the reference vector element in that row: that is, that the test vector contains extra, presumably spurious, peaks. These can be excluded from the warped vector by choosing the element of the test vector which gives the lowest distance in that row; this element is marked by an asterisk in Table 6.5. Similarly, where the path runs vertically down a column (e.g. column 3, Table 6.5), it indicates that the test vector is missing one or more peaks; the peak which is actually present is identified by choosing the element with the lowest distance in that column (marked with an asterisk), and a code indicating a "missing value" is inserted into the other locations in the test vector. This is unlike DTW, which allows the *duplication* of analysis frames where the test token does not contain frames found in the prototype.

Figure 6.12 shows how this is applied to the matrix of distances for the last two tokens mentioned above, with token D treated as the prototype. The peaks at 611 Hz and 3418 Hz have been excluded from the result, while "missing values" have been inserted in second and fourth positions, corresponding to the peaks at 1619 Hz and 2492 Hz in the prototype.

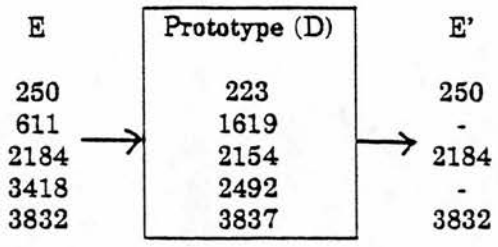

Figure 6.12 Effect of spectral peak warping on an analysis vector

It should also be noted that, though this method is referred to as peak "warping", the frequencies of the peaks themselves are unchanged, as is their order in the spectrum: only their relative alignment to the peaks of the prototype vector is altered.

6.5.2. The choice of a prototype

A method is needed for arriving at a suitable prototype. One possibility is the vector which produces the best overall alignment, and one way to measure this is to consider the *variability* of the parameters which make up the vectors. An obvious sign that the tokens are mis-aligned is that the values of each corresponding element vary widely: this is, in fact, what alerted us to the problem in the first place. One result of imposing a good alignment is, presumably, to reduce this variation, as tokens of the same underlying peak occupy the same places in each vector. One index of the success of a particular alignment, then, is the set of variances of the parameters, measured over the re-aligned vectors.

Table 6.6 shows the variances for the parameters of the set of aligned tokens produced when each token in turn is taken as the prototype. Missing values are omitted from the calculation of both means and variances in this table. The *average* variance gives a good indication of the success of each prototype, and can be used to select the corresponding vector as the prototype on which to base the warping of the remaining vectors in the database. In this set of tokens, profile A clearly gives the lowest average variance. An examination of Figure 6.11 suggests that this is quite a reasonable choice for a prototype, since the four peaks represented by token A are indeed the most consistent in their appearance.

The choice of a single, actually occurring token for the prototype, however, makes the result rather too dependent on the vagaries of the sample of utterances used. To make the algorithm more robust, a set of tokens can be chosen – the best five, for example – and their mean vector used as the prototype in place of any actually occurring vector. This reduces the chances of obtaining a deviant prototype, and thus of omitting useful data or unreasonably distorting

Parameter	Prototype				
	A	B	C	D	E
1	170.64	170.64	170.64	170.64	170.64
2	110.24	51529.00	3354.00	3354.00	51529.00
3	430.89	3354.00	110.24	110.24	110.24
4	523.60	110.24	430.89	430.89	0.00
5		523.60	523.60	5236.60	523.60
Mean	308.84	11137.50	917.87	917.87	10466.70

Table 6.6 Variances of aligned parameters for each prototype in turn

valid data.

6.6. Application of peak profile warping to the pole-zero analysis

The method outlined in the preceding section for selecting a prototype token and aligning the remaining tokens against it was applied to the pole and zero profiles obtained in section 6.3.

6.6.1. The choice of prototypes

There were several options for selecting a prototype. It seemed reasonable to assume that prototypes should be chosen separately for each sex, since the sexes can be expected to show certain differences in the number and distribution of spectral features. For the same reason, it was decided to find prototypes for each speaker separately: since we expect to find considerable differences among speakers, even of the same sex, the assumption of a shared underlying spectral pattern is invalid. Differences among the three vowel contexts were not so certain, but it was decided that vowel-specific prototypes should also be used.

There is obviously some danger that such data manipulation will introduce the very differences of sex, speaker and vowel context under investigation. However, it was felt that the variability of the raw data justified this approach, since without some manipulation any statistical analysis would be meaningless.

One common constraint which was imposed on all speakers was on the number of peaks and dips in the prototypes: it was decided to limit the search

for peak prototypes to those peak profiles having *six* peaks in the case of males and *five* peaks in the case of females; the search for dip prototypes was restricted to dip profiles having *two* dips in both sexes. These figures were chosen as a compromise between the desire not to exclude too many possibly useful spectral features, and the need to prevent the inclusion of spurious peaks. Both are above the modes of the distribution of the numbers of peaks and dips in the male and female speakers, but produce fairly large sets of candidate vectors for each speaker.

For each speaker, in each vowel context, two prototypes were selected: each six-peak (or five-peak) profile was used in turn to align the elements of all other peak profiles for that speaker, and the five profiles giving the lowest average variance over the six (five) parameters were averaged to give the prototype for that speaker; a similar procedure was used to choose a dip prototype. The two searches were done independently, and the two prototypes did not necessarily come from the same speech tokens. This process was repeated for each speaker separately.

6.6.2. Effects of realignment using prototypes

The chosen prototypes were then used to align each speaker's profiles, and these aligned profiles, all having six peaks (five for the females) and two dips (including any missing values) were used in all subsequent experiments in this Chapter.

Some idea of the effect of this realignment can be gained by comparing the variability of the unwarped peak and dip frequencies with that of their warped

counterparts. A suitable index of variability, which is independent of differences in scale, is the *coefficient of variation* (Spiegel 1961), defined as the ratio of the standard deviation of a distribution to the mean of that distribution; this ratio is usually expressed as a percentage. Figures 6.13 to 6.16 compare the coefficients of variation for the eight parameters – the first six peaks and first two dips of the raw data and the six (five) peaks and two dips of the warped data – for male and female speakers. For the females, no warped equivalent of the sixth peak exists.

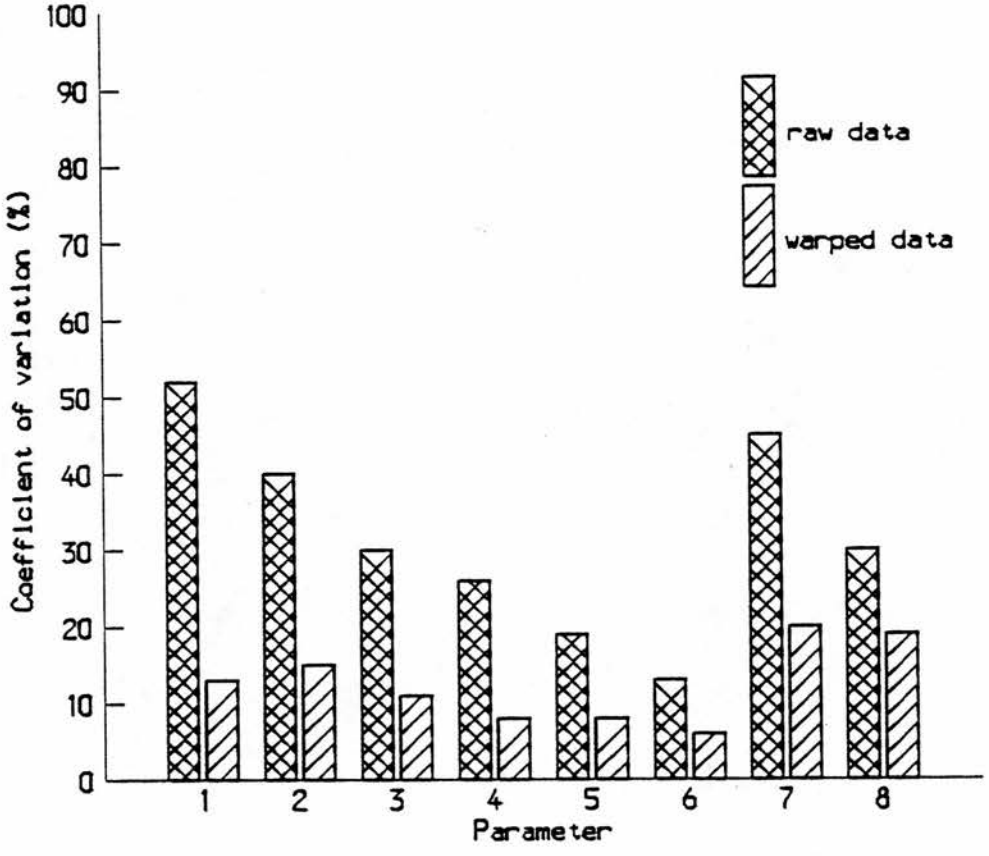

Figure 6.13 Mean coefficients of variation of raw and warped parameters for 15 male speakers

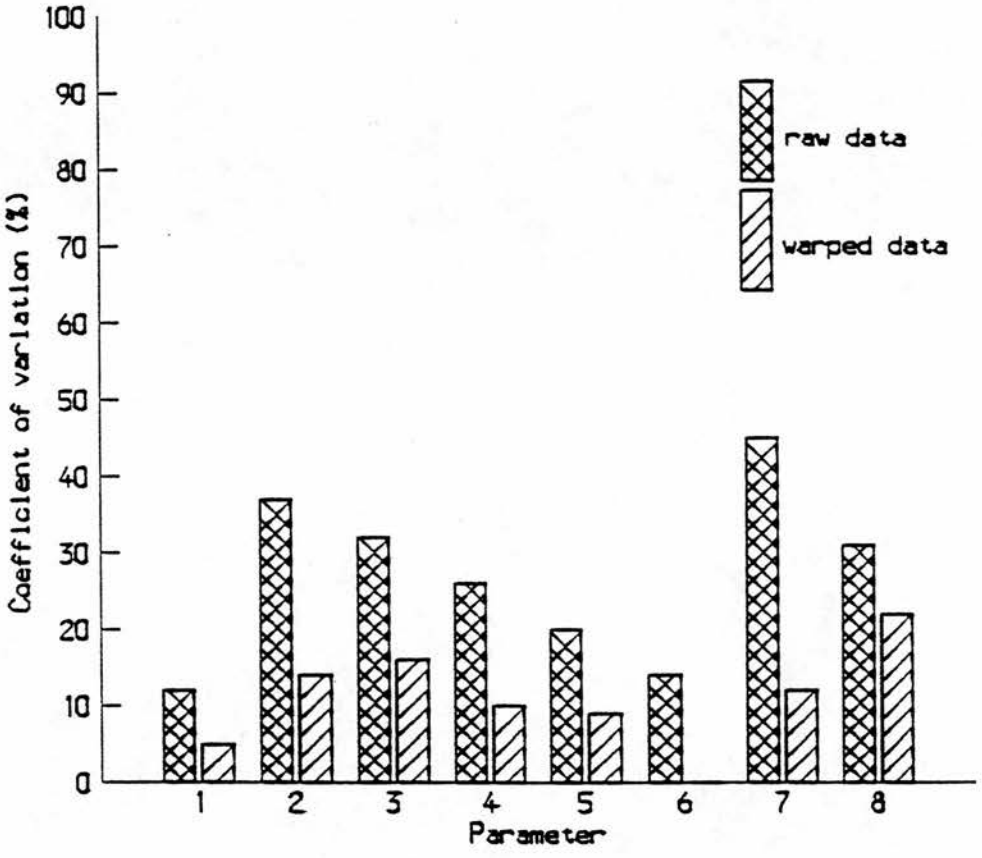

Figure 6.14 Mean coefficients of variation of raw and warped parameters for 15 female speakers

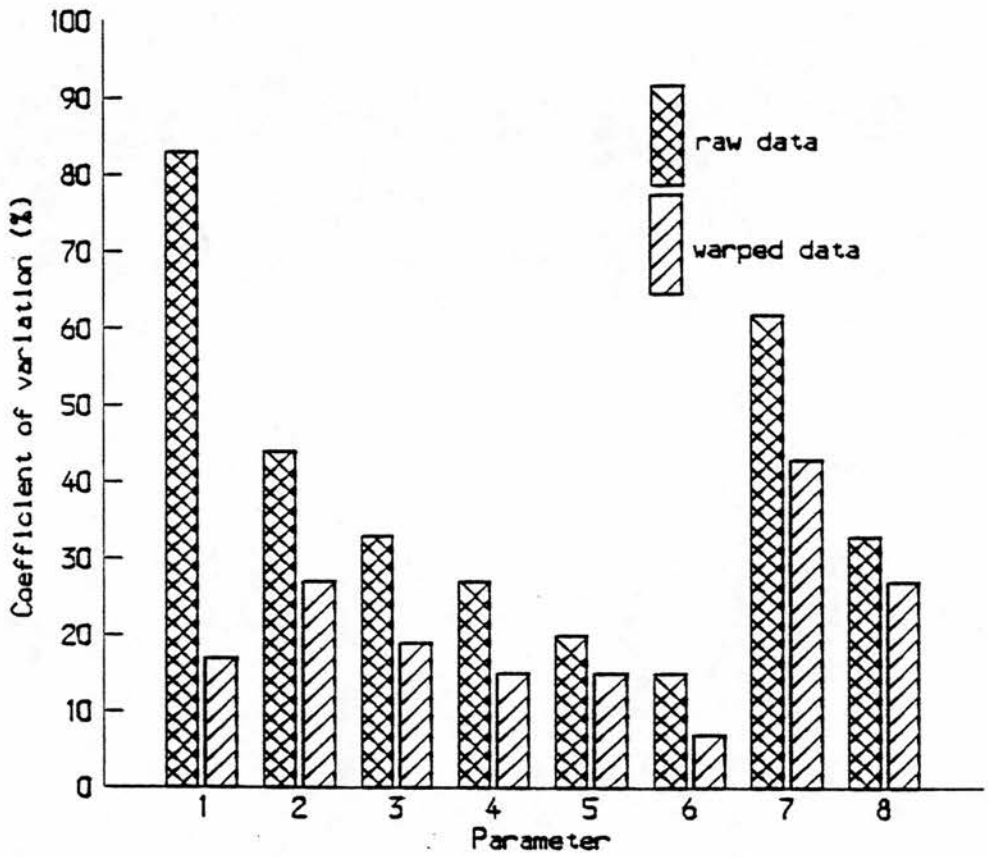

Figure 6.15 Pooled coefficients of variation of raw and warped parameters for 15 male speakers

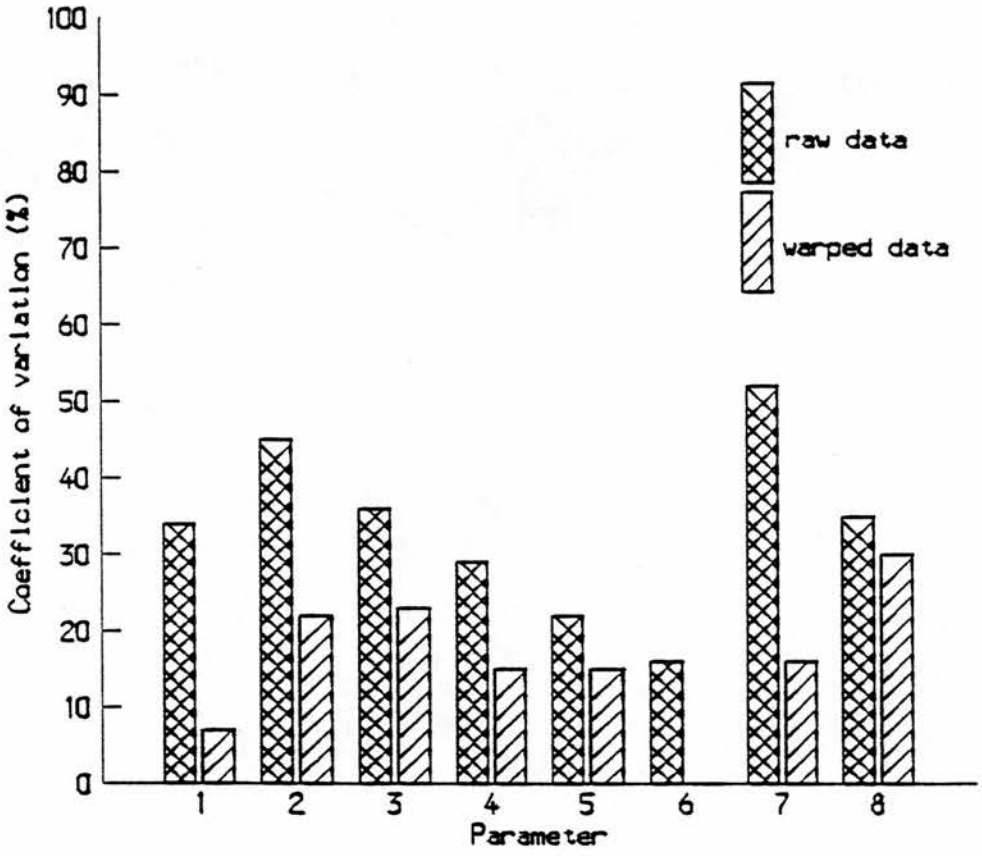

Figure 6.16 Pooled coefficients of variation of raw and warped parameters for 15 female speakers

In Figures 6.13 and 6.14 the coefficient of variation was derived for each speaker separately and then averaged over all speakers in the group. There is a clear and consistent reduction in the variability of all eight parameters, as might be expected from the method used to realign the data. In Figures 6.15 and 6.16 the coefficient of variation was calculated for all speakers together, so that between speaker variation was also included. It can be seen that the reduction in variability is maintained, suggesting that the procedure does not enhance between-speaker differences despite the fact that speaker-dependent prototypes are used.

6.7. Sex differences

The warped data for the thirty speakers were analysed to see if the expected differences between the sexes were present. Female speakers generally have a shorter vocal tract length, and also a shorter and narrower nasal tract (Bjuggren and Fant 1964). We might expect the fundamental resonance of their nasal-pharyngeal tube to be higher than that of male speakers, with a concomitant increase in frequency of all higher resonance peaks.

Figures 6.17 and 6.18 present histograms for the eight parameters for male and female groups respectively; the female speakers have no values for peak 6, since only five values were used in their prototypes. Statistical summaries (means and standard deviations) are given in Table 6.7. The mean values are presented in Figure 6.19.

Group	P1	P2	P3	P4	P5	P6	Z1	Z2
Males	266.8	1212.0	1894.0	2483.0	3269.9	4132.3	1057.7	2768.6
	(46.7)	(329.8)	(355.8)	(363.7)	(478.9)	(304.9)	(459.1)	(753.3)
Females	268.1	1241.3	1907.8	2782.5	3964.4	-	887.1	2322.7
	(19.1)	(273.7)	(433.4)	(407.5)	(597.9)	(-)	(139.7)	(704.3)

Table 6.7 Parameter means and standard deviations (warped data) for male and female speakers

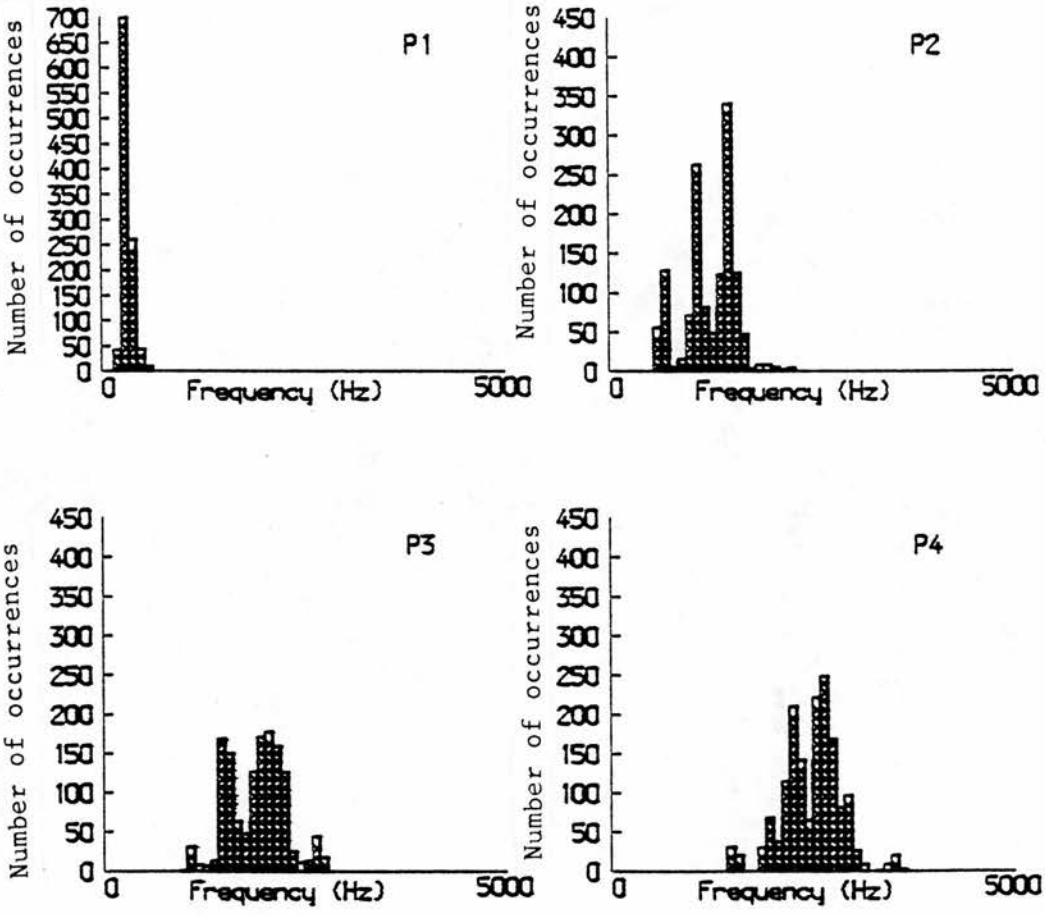

Figure 6.17 Pole and zero frequency distributions (warped data) for 15 male speakers

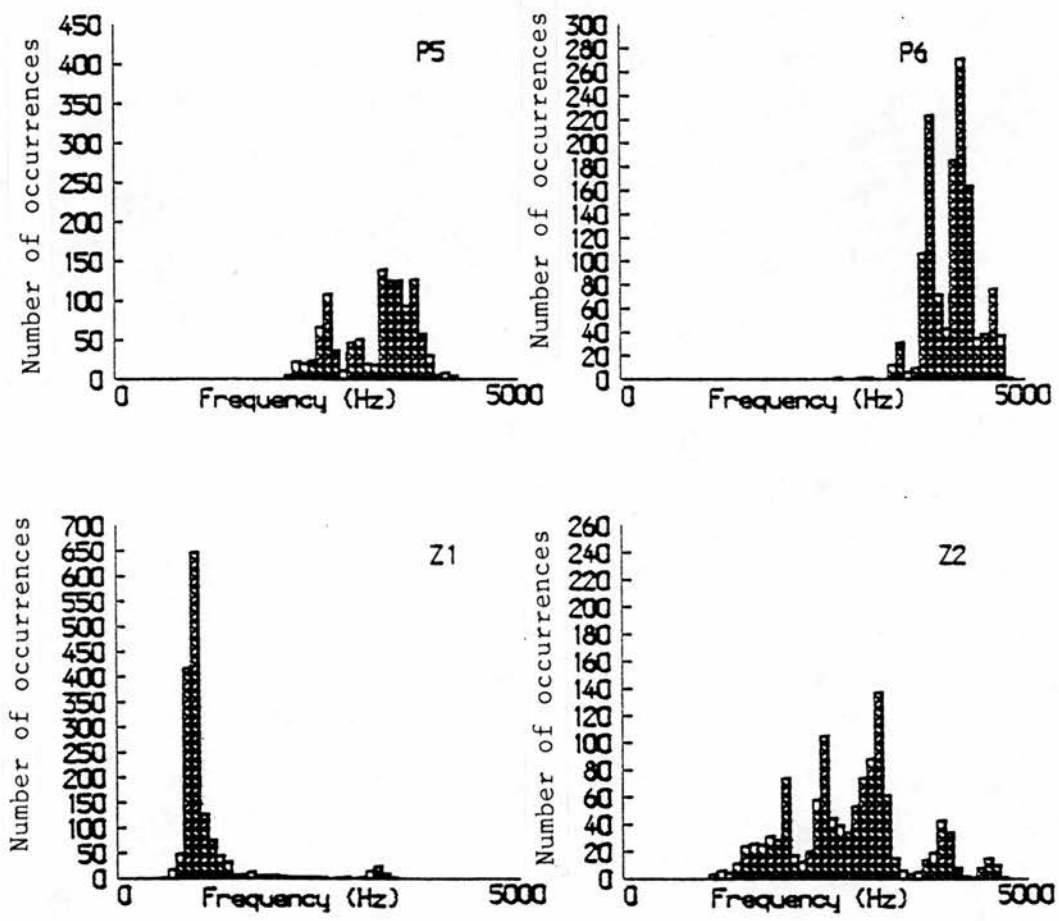

Figure 6.17 Pole and zero frequency distributions (warped data) for 15 male speakers

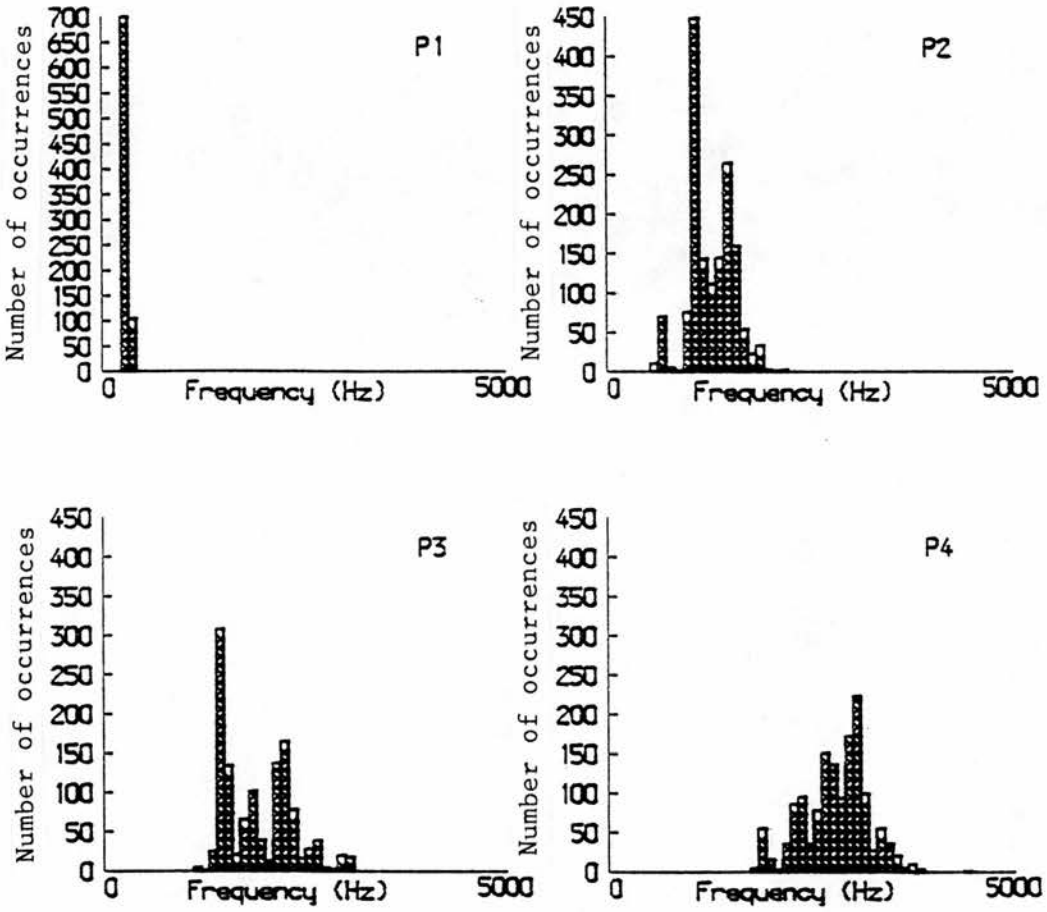

Figure 6.18 Pole and zero frequency distributions (warped data) for 15 female speakers

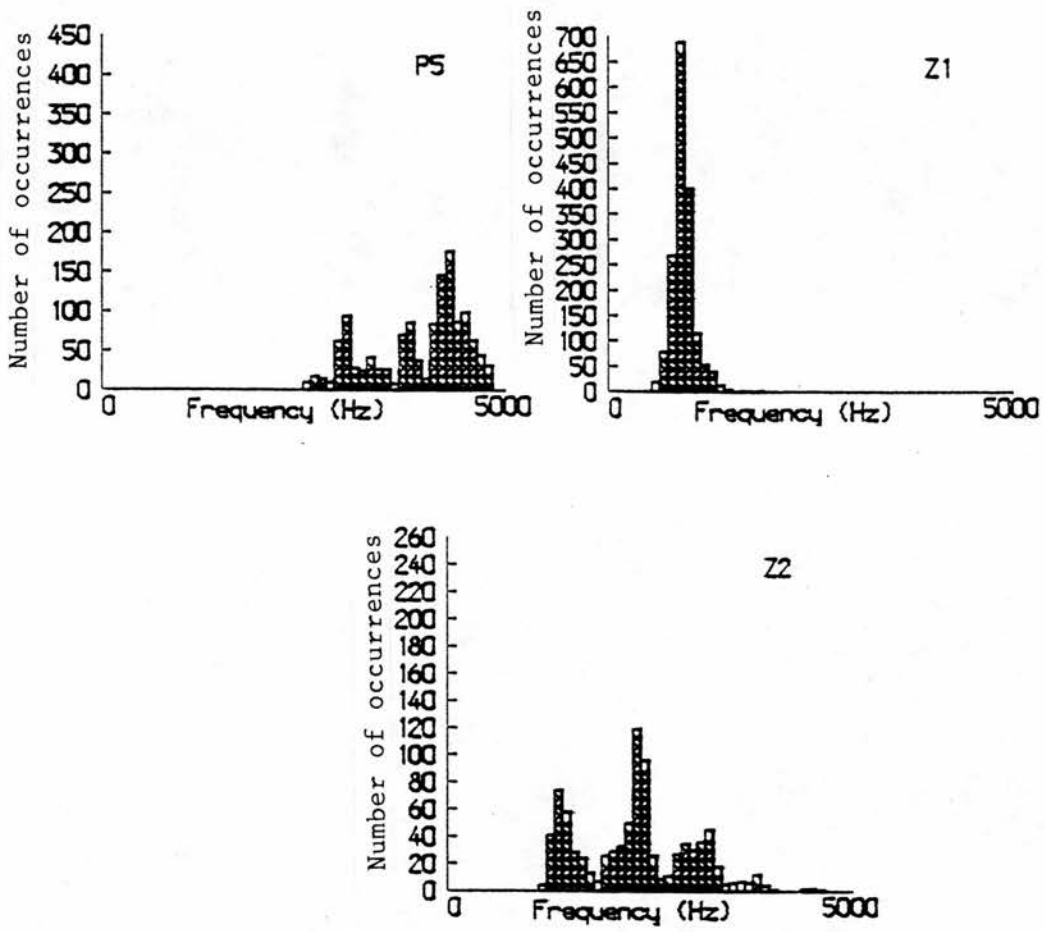

Figure 6.18 Pole and zero frequency distributions (warped data) for 15 female speakers

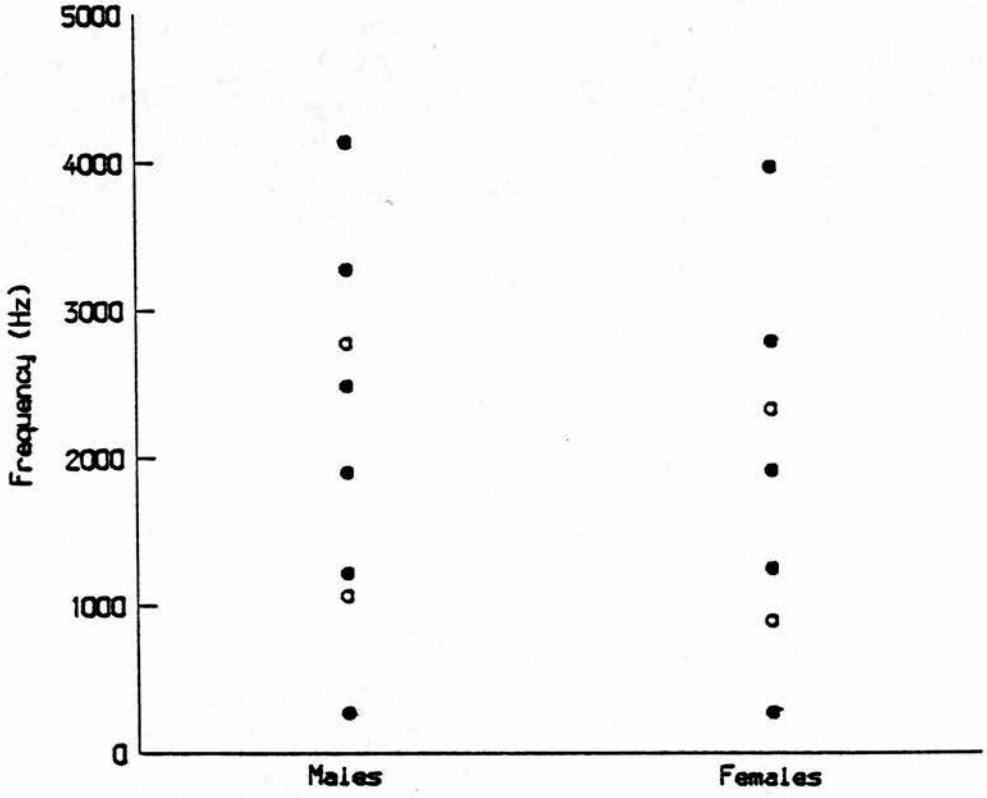

Figure 6.19 Mean pole and zero frequencies for 15 male and 15 female speakers (poles •, zeros o)

The results of t-tests for differences in the group means of the first five peaks and the two zeros are presented in Table 6.8. In this table, as in all subsequent analyses, values which are significant at a probability level of 0.001 or lower are marked with an asterisk. Values of t are significant only for the higher peaks (4 and 5) and the two zero parameters. The direction of the differences between corresponding peaks is consistent with the hypothesis that female speakers have higher values, but the opposite is true in the case of the zeros, the female means being significantly lower than those of the males. Examination of the histograms suggests that the reason for our failure to find the expected sex difference in the lower peaks is that the two groups show too much internal variability, the male speakers having a particularly wide range. A similar explanation can be offered for the reversal of the expected difference between the male and female zero frequencies: in the case of Z1, the male speakers show a small number of outlying values which should probably have been identified with Z2, and which raise the mean value to above that of the

Parameter	t
P1	-1.14
P2	-2.58
P3	-0.86
P4	-21.28*
P5	-31.81*
P6	-
Z1	13.82*
Z2	13.77*

Table 6.8 Values of t for difference between male and female group means

females, even though the main part of the distribution lies at a lower frequency.

6.8. Effects of vowel context (coarticulation)

It was noted in Chapter Three that there is little published information on the extent and nature of vowel context effects in the velar nasal, other than that the velar is less susceptible to coarticulatory changes than the alveolar and bilabial nasals. No information at all has been published on changes in the zero (anti-resonance) frequencies.

It is possible to predict some of the changes which might occur in the spectrum of the velar nasal for some vowel contexts. For example, it is known that the tongue makes contact with the soft palate further forward in the context of high front vowels such as /ii/, at least in oral stops (e.g. Ohman 1967, Gimson 1970); if this were the case in velar nasal stops too, the lengthening of the side-chamber behind the oral constriction should lead to a lowering of any anti-resonances produced in the spectrum, while the enlargement of the entire nasal-pharyngeal-oral system should cause a lowering of all the resonances which that system produces. We might also expect the velar contact to be made further back in the case of the /uh/-/u/ vowel context, causing a rise in both resonance and anti-resonance frequencies (though in the case of speakers with /u/ rather than /uh/, the lip-rounding might counteract this tendency). The "neutral" case would be that of the vowel /a/.

Figure 6.20 and 6.21 show the mean peak and dip frequencies for the three vowel contexts, pooled over male speakers and female speakers respectively.

These data are also presented in Table 6.9. The differences between the vowel contexts appear rather small, and are not completely consistent with the tentative predictions made above. For example, while the male speakers' first zero frequency behaves as expected, being lower for the /i/ context than for /a/ and /uh/, the opposite is true for the female speakers, with the /i/ context showing a *higher* mean value.

Males								
Type	P1	P2	P3	P4	P5	P6	Z1	Z2
ing	265.9	1217.3	1857.1	2467.9	3176.0	4091.0	975.8	2763.7
	(42.1)	(296.3)	(269.4)	(254.1)	(445.0)	(286.2)	(214.9)	(734.3)
ang	269.0	1230.8	1924.8	2497.8	3462.3	4150.6	1111.5	2729.5
	(52.3)	(324.7)	(341.6)	(330.2)	(410.7)	(305.2)	(552.9)	(740.5)
ung	265.5	1187.3	1899.7	2483.8	3206.8	4159.2	1091.0	2814.4
	(45.1)	(364.2)	(440.8)	(470.7)	(508.7)	(320.0)	(530.2)	(786.8)
Females								
Type	P1	P2	P3	P4	P5	P6	Z1	Z2
ing	269.8	1307.6	2105.2	2791.3	3946.5	-	908.4	2224.8
	(18.3)	(262.8)	(391.8)	(353.2)	(677.8)	(-)	(117.3)	(659.2)
ang	266.4	1227.5	1859.0	2759.2	4023.9	-	896.2	2467.6
	(19.6)	(282.3)	(382.2)	(371.3)	(553.0)	(-)	(164.2)	(731.0)
ung	267.9	1177.2	1750.9	2796.3	3924.6	-	853.4	2294.1
	(19.2)	(259.4)	(450.3)	(481.2)	(548.9)	(-)	(128.7)	(705.6)

Table 6.9 Parameter means and standard deviations (warped data) by vowel context

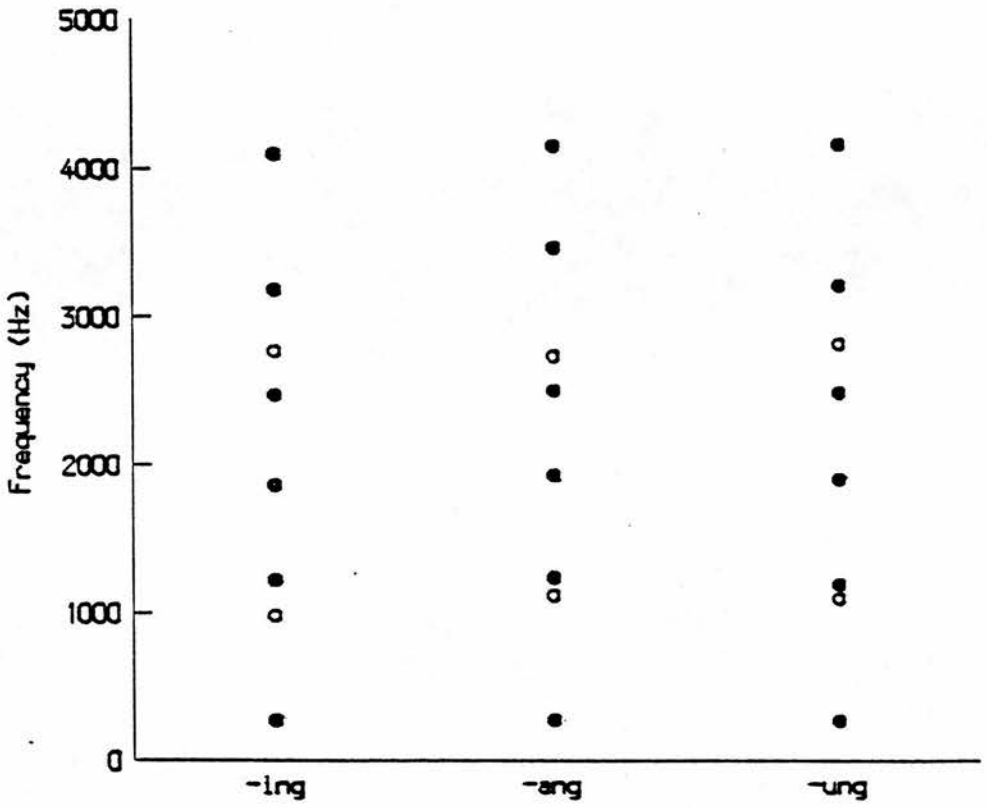

Figure 6.20 Mean pole and zero frequencies by vowel context for 15 male speakers (poles •, zeros ○)

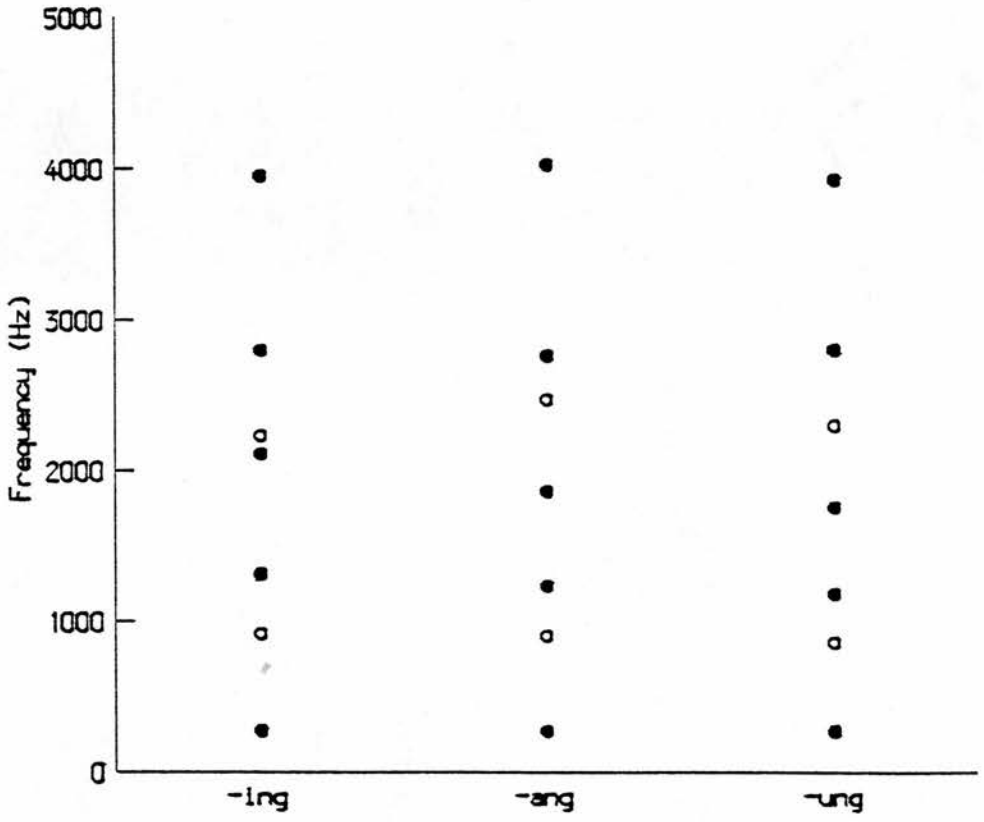

Figure 6.21 Mean pole and zero frequencies by vowel context for 15 female speakers (poles ●, zeros ○)

A one-way analysis of variance confirmed that in most cases the differences are insignificant (Table 6.10).

A possible explanation for this result is that any difference there is among vowel contexts is swamped by the differences between speakers. Figures 6.22 (males) and 6.23 (females) show the effect of vowel context on the mean value of each of the eight parameters for each speaker. These figures indicate that there are indeed large differences between speakers in the effects of vowel context. It is not possible, however, to distinguish any clear pattern consistent with the tentative hypotheses presented above.

Parameter	Males	Females
P1	1.13	5.37
P2	2.03	31.79*
P3	4.25	78.91*
P4	0.9	1.18
P5	41.12*	3.29
P6	6.75	-
P7	13.09*	24.1*
P8	1.18	8.91*

Table 6.10 F-ratios showing the effect of vowel context (warped data)

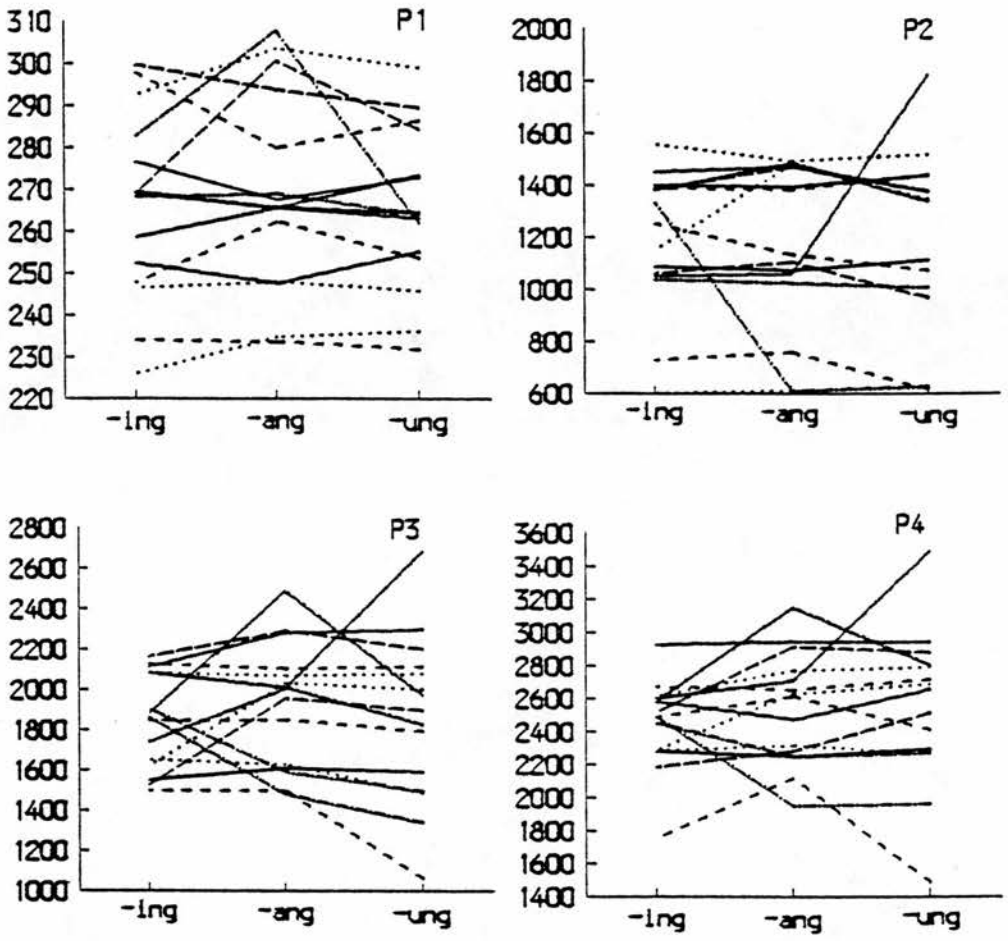

Figure 6.22 Mean frequencies (Hz) for each male speaker according to vowel context

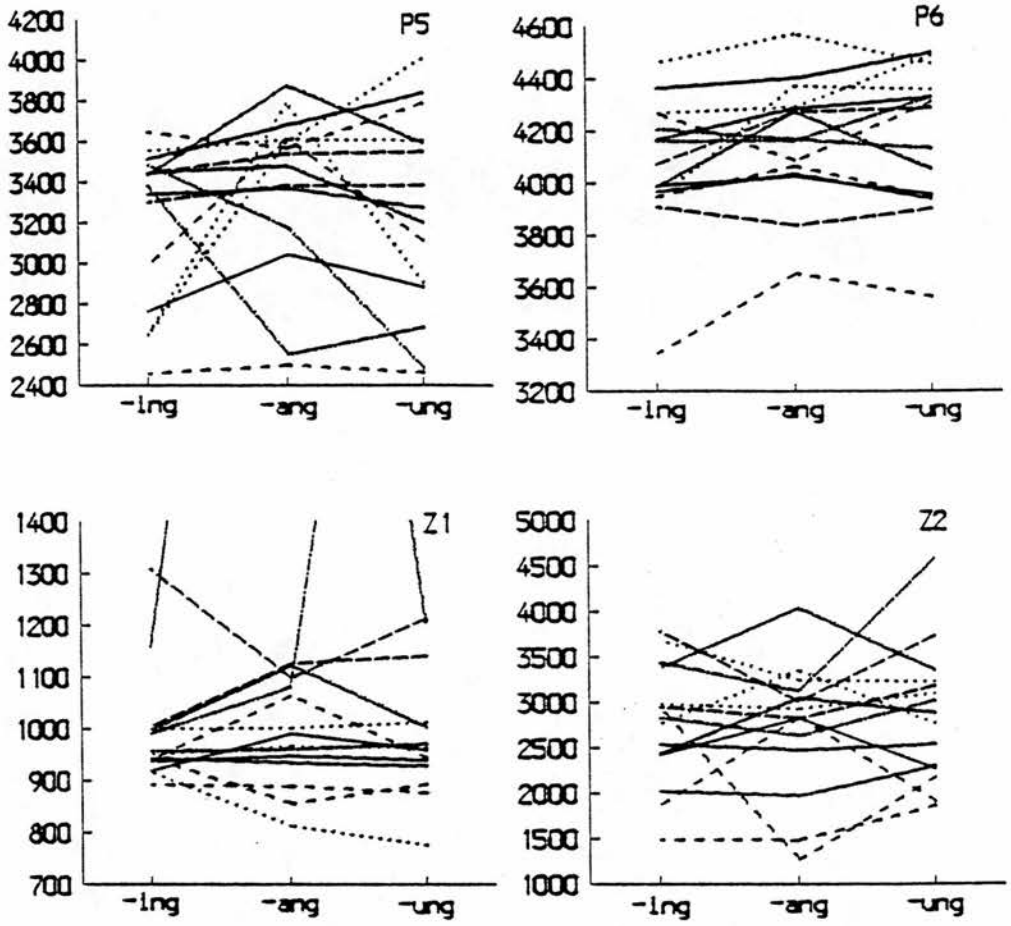

Figure 6.22 Mean frequencies (Hz) for each male speaker according to vowel context

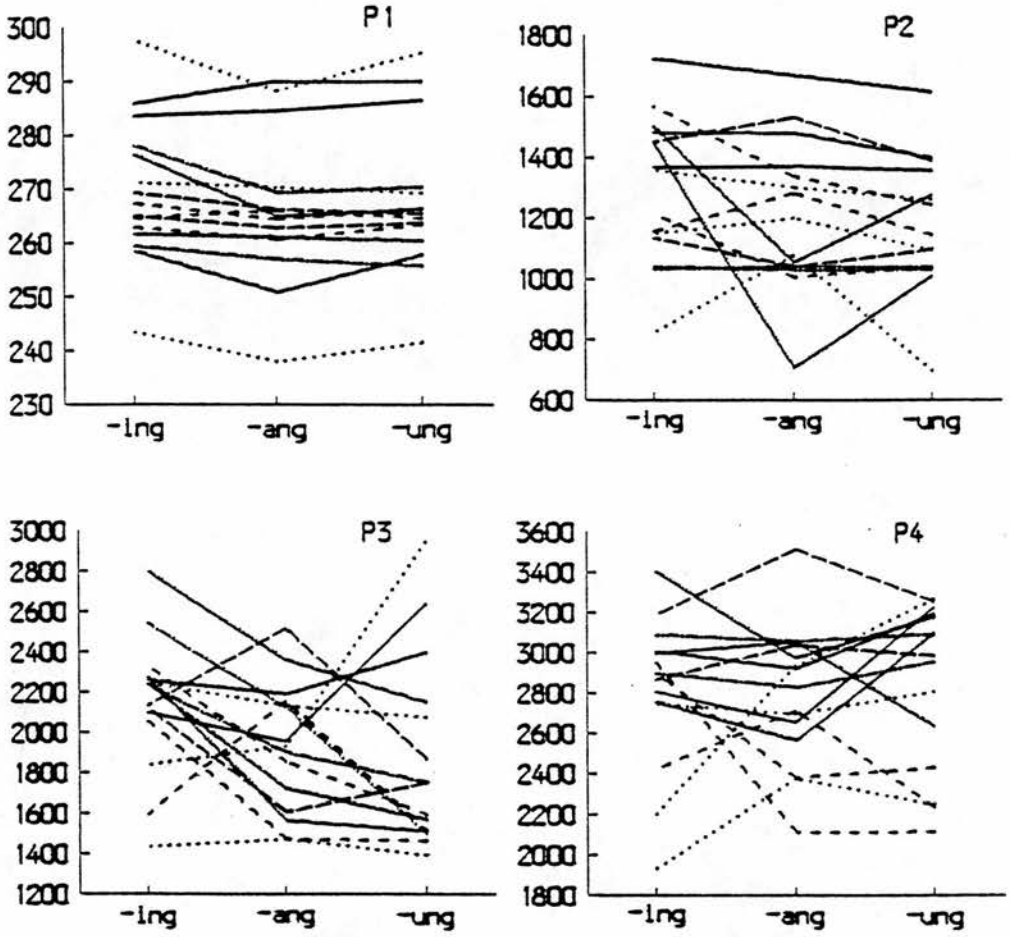

Figure 6.23 Mean frequencies (Hz) for each female speaker according to vowel context

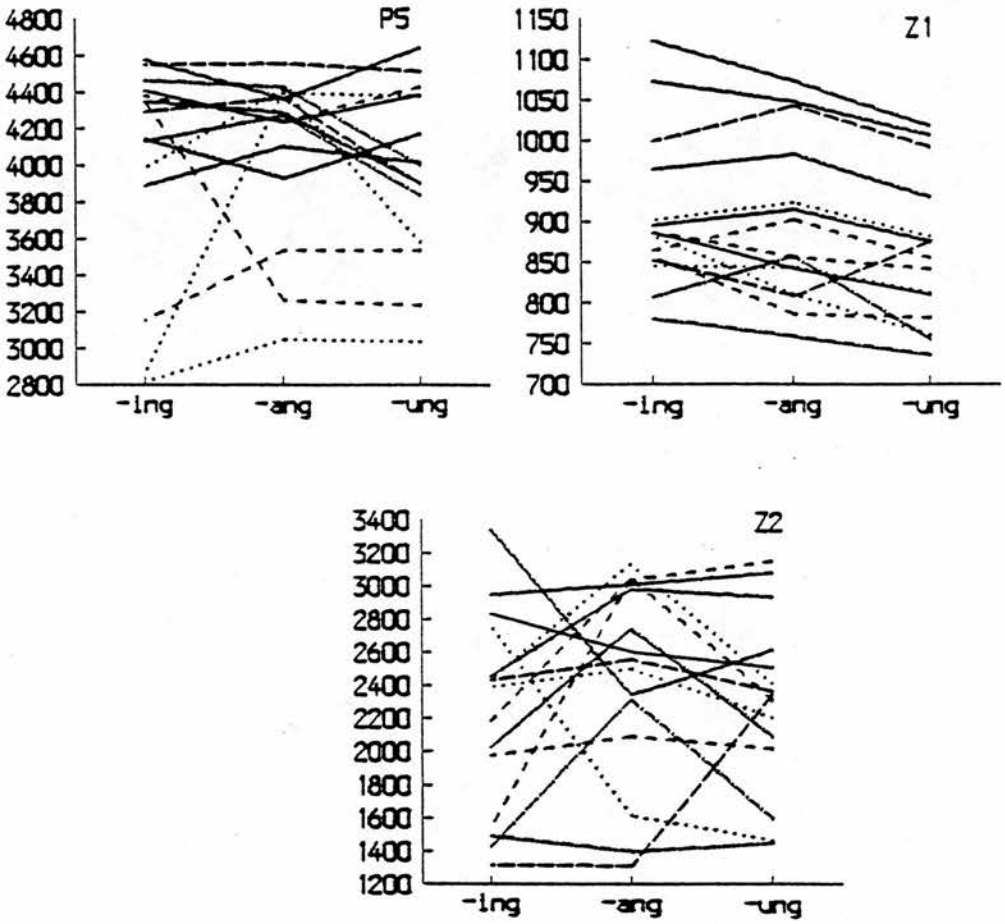

Figure 6.23 Mean frequencies (Hz) for each female speaker according to vowel context

Thus while the data show evidence for some significant vowel context effects, and the effect of vowel context cannot yet be discounted, no clear pattern is found. It seems that even within the performance of a single speaker values of the eight parameters show too much variability. A better approach to studying vowel-context effects themselves might be to use a smaller group of speakers and to impose greater control over the phonetic environment, possibly restricting the recordings to a single session.

6.9. Within and between speaker variability

A central theme of this thesis is the extent of the variability within the speech of each speaker, and the degree of difference found between speakers: low within-speaker variability and relatively high between-speaker variability are crucial for success in automatic speaker verification (see Chapter Two, 2.4). The measurement of within-speaker variability in this chapter has been approached using the analysis of variance, rather than by attempting to track changes in the eight parameters for each speaker over the eight recording sessions. It was considered that such an approach based on the time at which the recordings were made would be unlikely to show any trends of interest, since any variation over this time would presumably be due to random fluctuations in speakers' performance, or to short-term health changes which could not easily be tracked, rather than to any structural change in the speaker.

Some idea of the nature of the differences between speakers, and the amount of variability within speakers' performance may be obtained from an inspection of the distribution of their peak and dip frequencies. These

distributions are summarised for each speaker in Figures 6.24 (males) and 6.25 (females), and in the accompanying Tables 6.11 and 6.12, which give the mean frequency for each peak or dip, together with the standard deviation.

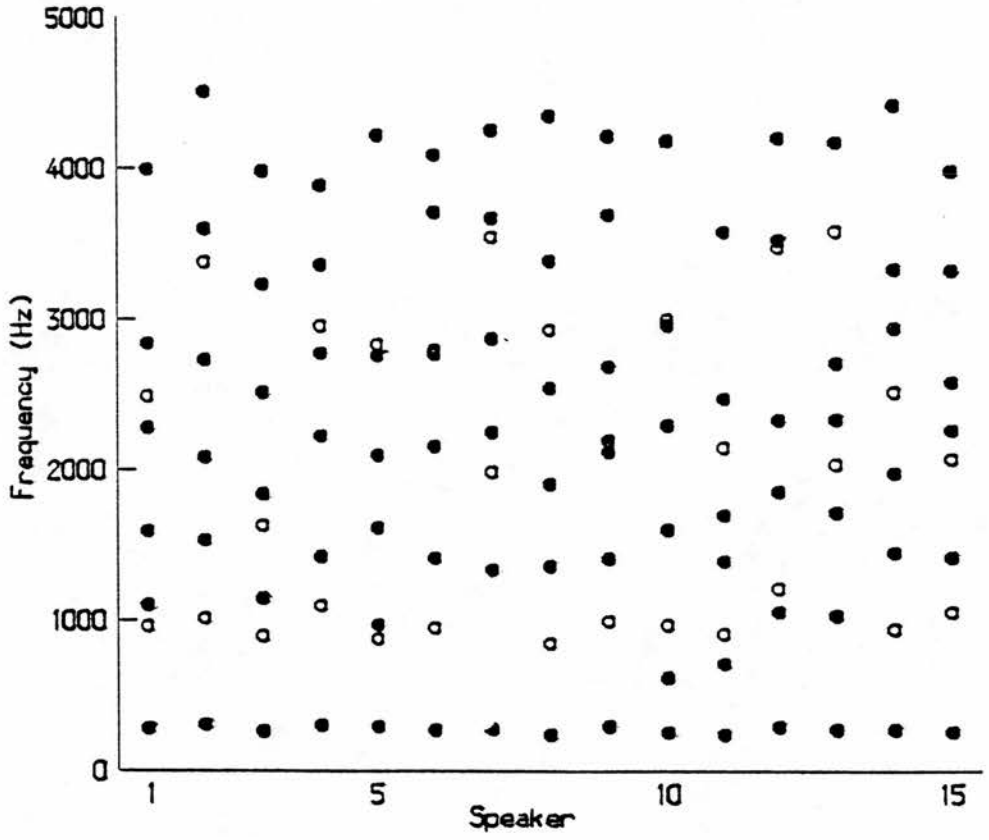

Figure 6.24 Mean pole and zero frequencies for each male speaker (poles •, zeros o)

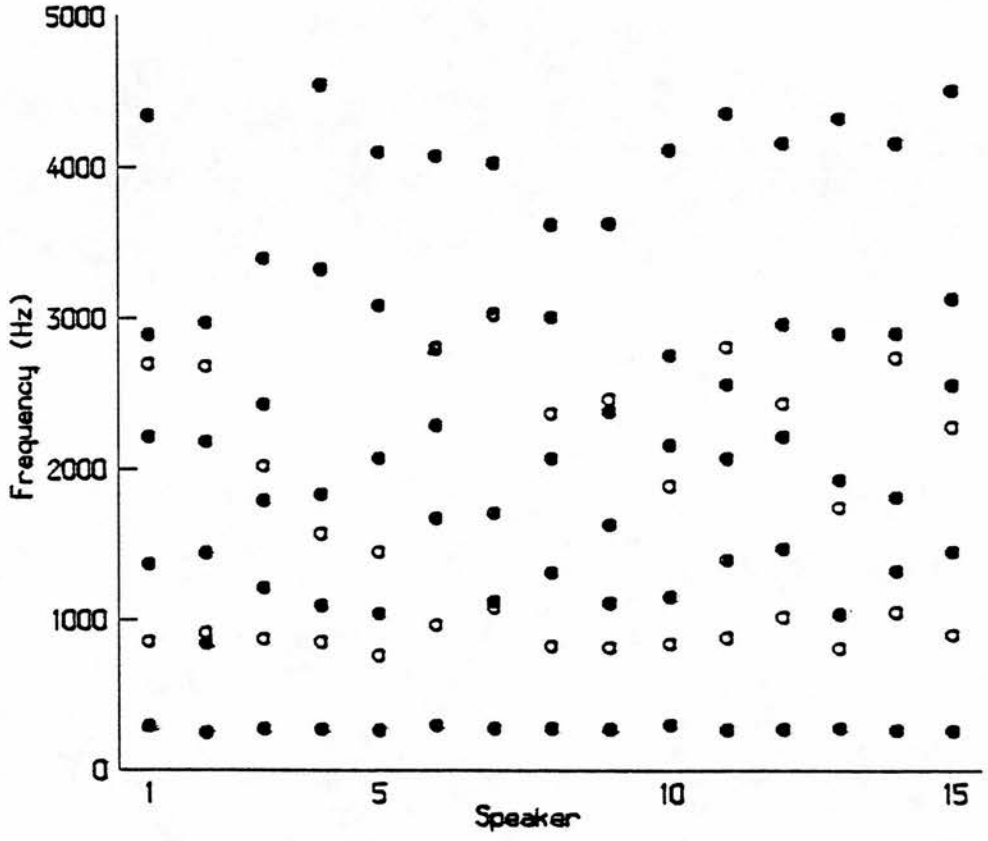

Figure 6.25 Mean pole and zero frequencies for each female speaker (poles •, zeros o)

It is clear from these figures that there are large differences among speakers, particularly in the frequencies of the second dip and of many of the higher peaks. The first peak is probably the most consistent across speakers, but in

Speaker	P1	P2	P3	P4	P5	P6	Z1	Z2
1	272.3	1091.8	1583.6	2270.9	2830.4	3985.9	951.4	2479.5
	(59.9)	(120.6)	(152.6)	(179.3)	(244.9)	(302.4)	(191.8)	(404.7)
2	298.3	1525.2	2075.4	2721.9	3591.6	4500.4	1004.9	3370.1
	(33.4)	(141.4)	(51.4)	(111.1)	(114.7)	(160.0)	(153.1)	(448.4)
3	254.6	1137.9	1832.5	2504.9	3223.8	3974.2	886.8	1622.8
	(31.1)	(149.3)	(164.9)	(245.5)	(309.0)	(261.2)	(47.4)	(402.7)
4	294.2	1415.1	2217.7	2767.3	3353.6	3880.4	1090.1	2948.9
	(41.2)	(216.1)	(117.1)	(218.7)	(105.4)	(173.9)	(234.2)	(469.3)
5	284.7	868.8	1607.3	2088.9	2753.0	4213.6	961.8	2825.2
	(116.6)	(350.2)	(282.8)	(255.3)	(494.1)	(246.6)	(106.4)	(440.4)
6	262.9	1405.3	2148.5	2760.6	3703.0	4082.6	941.1	2786.2
	(15.2)	(159.1)	(342.2)	(252.1)	(269.7)	(224.7)	(73.4)	(498.6)
7	269.3	1328.0	2241.6	2862.9	3663.5	4247.7	1977.8	3539.0
	(19.3)	(422.4)	(461.2)	(376.1)	(194.4)	(157.9)	(1019.3)	(495.1)
8	232.2	1352.3	1897.0	2534.4	3378.8	4339.7	838.3	2920.1
	(34.2)	(224.1)	(252.1)	(209.2)	(665.7)	(177.1)	(135.0)	(626.0)
9	288.1	1402.1	2114.9	2677.4	3686.5	4207.4	987.6	2186.3
	(56.2)	(150.7)	(104.8)	(115.6)	(182.3)	(204.3)	(179.6)	(635.4)
10	246.7	610.3	1594.1	2289.1	2953.3	4178.9	959.9	2989.6
	(10.0)	(38.1)	(160.0)	(65.9)	(453.6)	(329.0)	(41.6)	(341.6)
11	233.1	701.9	1382.8	1691.5	2465.4	3572.2	900.0	2140.0
	(9.2)	(184.0)	(197.4)	(291.9)	(168.0)	(340.3)	(144.0)	(742.2)
12	284.1	1046.9	1845.4	2323.2	3519.3	4195.7	1204.1	3469.7
	(36.8)	(164.0)	(206.4)	(202.8)	(173.7)	(264.5)	(484.0)	(588.9)
13	267.1	1023.7	1709.8	2327.1	2702.1	4167.0	2028.4	3579.0
	(24.1)	(72.5)	(269.8)	(144.2)	(237.6)	(160.2)	(1097.5)	(727.1)
14	265.9	1444.0	1971.1	2934.1	3327.4	4415.6	934.9	2510.4
	(17.8)	(111.5)	(186.5)	(94.4)	(227.3)	(176.5)	(39.3)	(202.3)
15	251.6	1411.4	2255.7	2575.7	3317.8	3978.4	1048.3	2064.8
	(26.6)	(131.8)	(146.4)	(164.1)	(138.4)	(199.9)	(139.8)	(424.0)

Table 6.11 Means and standard deviations of peak and dip frequencies for 15 male speakers

Speaker	P1	P2	P3	P4	P5	Z1	Z2
1	284.8	1362.6	2207.4	2883.6	4336.5	846.8	2688.4
	(13.7)	(123.1)	(390.6)	(318.2)	(229.6)	(99.5)	(508.9)
2	240.9	836.1	1436.0	2175.8	2963.1	904.4	2674.1
	(13.3)	(227.8)	(88.6)	(344.1)	(199.8)	(124.8)	(563.8)
3	265.1	1206.2	1785.1	2421.8	3389.4	863.7	2014.8
	(9.0)	(150.1)	(371.9)	(282.6)	(332.5)	(79.4)	(293.9)
4	263.8	1088.4	1826.8	3319.8	4541.3	846.0	1564.2
	(7.2)	(136.2)	(302.1)	(207.3)	(150.3)	(86.0)	(591.2)
5	257.4	1035.8	2065.4	3077.2	4096.5	757.0	1442.8
	(12.1)	(33.1)	(525.8)	(265.8)	(370.9)	(92.9)	(285.3)
6	288.6	1667.2	2283.6	2791.8	4070.1	959.2	2801.9
	(25.3)	(176.8)	(180.9)	(240.7)	(243.1)	(130.0)	(556.3)
7	269.3	1115.1	1702.5	3024.1	4025.1	1076.1	3017.3
	(12.9)	(364.7)	(348.5)	(313.7)	(316.5)	(151.8)	(362.3)
8	270.3	1307.7	2065.4	3001.2	3615.8	817.9	2363.5
	(8.6)	(115.6)	(437.1)	(387.4)	(784.3)	(73.5)	(246.1)
9	266.0	1104.9	1626.8	2378.2	3625.0	808.4	2459.3
	(9.5)	(177.2)	(279.9)	(432.0)	(596.3)	(108.3)	(566.2)
10	293.7	1146.8	2156.1	2750.7	4118.2	834.0	1884.3
	(17.4)	(106.1)	(177.5)	(128.9)	(472.7)	(57.2)	(628.0)
11	262.3	1392.8	2067.4	2559.1	4363.6	875.1	2805.7
	(10.3)	(227.0)	(351.6)	(293.8)	(290.7)	(157.7)	(768.8)
12	267.0	1467.9	2212.2	2955.8	4165.1	1012.2	2430.0
	(5.7)	(126.1)	(308.2)	(233.9)	(389.6)	(101.9)	(188.1)
13	272.6	1031.8	1924.9	2893.1	4328.7	804.3	1740.8
	(15.9)	(39.0)	(337.9)	(220.2)	(337.4)	(141.8)	(584.9)
14	261.1	1320.1	1809.9	2896.1	4166.5	1045.3	2735.3
	(11.6)	(328.0)	(353.4)	(299.9)	(342.7)	(109.0)	(633.4)
15	255.7	1447.6	2552.6	3126.1	4517.0	894.8	2276.8
	(19.5)	(136.0)	(383.6)	(289.5)	(312.0)	(64.6)	(549.4)

Table 6.12 Means and standard deviations of peak and dip frequencies for 15 female speakers

other places the mismatch between speakers is obvious: for example, some show a second peak below 1000 Hz, but for most this peak is missing. The application of peak profile warping may have contributed to some of these

differences, since separate prototypes were used for each speaker, but these were chosen in a way that reflected the general pattern of data for each speaker, so that differences between speakers can be assumed to be largely genuine.

The extent of within-speaker variability can be gauged from the standard deviation values of the eight parameters (given in Tables 6.11 and 6.12). The first peak is extremely consistent *within* speakers' performance as well as across speakers, standard deviations being generally less than 10% of the mean values. No clear pattern emerges from examination of the remaining parameters, however.

A more useful picture of the variability is obtained using the analysis of variance, relating within-speaker variation to between-speaker variation in a single index (the F-ratio) for each parameter. Separate analyses were carried out on the male and female speakers. Bearing in mind the effects of vowel context revealed in Section 6.8, the effects of speaker identity and vowel context were considered simultaneously, along with the interaction between them, in a two-way analysis of variance. The results of this analysis are shown in Table 6.13.

This analysis confirms that the major factor influencing the variability of the parameters is the identity of the speaker: that is, between speaker variability greatly exceeds the variability within speakers, and the variability due to vowel context. Only in one case (P3 for the female speakers) does the effect of vowel context exceed that of speaker identity. This result suggests, however,

Parameter		Males	Females
P1	Speaker	31.89*	132.67*
	Vowel	1.95	9.85*
	Interaction	1.67	1.62
P2	Speaker	310.5*	242.53*
	Vowel	2.02	99.48*
	Interaction	33.4*	29.13*
P3	Speaker	331.19*	158.18*
	Vowel	37.9*	183.89*
	Interaction	70.44*	62.65*
P4	Speaker	486.54*	171.44*
	Vowel	74.75*	8.79*
	Interaction	82.47*	42.96*
P5	Speaker	292.32*	260.87*
	Vowel	46.74*	19.03*
	Interaction	57.05*	52.1*
P6	Speaker	85.13*	-
	Vowel	29.88*	-
	Interaction	6.44*	-
Z1	Speaker	102.22*	82.64*
	Vowel	46.73*	27.06*
	Interaction	71.39*	2.25*
Z2	Speaker	166.33*	83.78*
	Vowel	8.09*	17.25*
	Interaction	20.69*	19.72*

Table 6.13 F-ratios (speaker by vowel) for poles and zeros

that vowel context should indeed be taken into consideration in the use of velar nasal features for speaker verification (as attempted in Chapter Seven, 7.4).

Table 6.14 shows the parameters ranked by their F-ratio for speaker for the male and female speakers. These data indicate that it is the peaks in the middle frequency range which show the greatest between-speaker variability and within-speaker consistency, for both males and females. This seems to agree with a comment made by Hattori and co-workers, that "these higher

Parameter	Males	Females
P1	8	5
P2	3	2
P3	2	4
P4	1	3
P5	4	1
P6	7	-
Z1	6	7
Z2	5	6

Table 6.14 Ranking by F-ratio (speaker) of eight warped parameters

modes (of both resonance and anti-resonance) would differ greatly from person to person" (1958: 271). The low rank of the first peak might be expected, given the evidence of Figure 6.24, but the poor ranking of the two zero parameters would not be predicted from the distribution of the speaker means, and can perhaps only be explained by poor within-speaker consistency (see Table 6.8). The effectiveness of all eight parameters will be explored further in the next chapter.

6.10. Discussion and summary

This chapter has examined the variability of the velar nasal spectrum using the cepstral decomposition technique to obtain estimates of spectral peaks (maxima of the all-pole response) and spectral dips (minima of the all-zero response) in a relatively large and heterogeneous group of speakers. The difficulties involved in obtaining peak and dip frequencies for the velar nasal proved to be considerable. Wide variations in the *number* of peaks and dips emerged. Some of these variations could result from differences between speak-

ers in the bandwidths of their resonances, since a fixed limit of 500 Hz was imposed for all speakers in the search for features (6.3.1). However, it is likely that these differences are genuine, given the potential for variability noted in Chapter Three (3.7).

This variability makes accurate assessment of spectral differences between speakers – and of spectral variation within speakers – rather difficult, since comparisons between elements which clearly belong to different vocal tract resonances give a false picture. It was therefore necessary to realign the peaks and dips located in the pole-zero spectra. A method of *peak profile warping* based on the notion of a prototypical resonance pattern or set of patterns for each speaker was therefore introduced, to impose an alignment on each speaker's vectors. This step was essential for any statistical analysis to take place, and appeared not to enhance between-speaker differences (6.6.2), even though speaker-dependent prototypes were used.

In the analyses made on the warped peaks and dip profiles, it was found that, while some parameters showed the expected patterns attributable to sex and vowel context effects, others did not. Part of the reason for the lack of clear trends appears to be the extent of the differences between speakers. These differences were shown to exceed the variation within speakers' performance, more so in the case of peak features than of dips. It is possible that the spectral dips observed relate more to oral cavity anti-resonance, for example, than to fixed anatomical features such as the paranasal sinuses.

Only one other study has considered the variability of the velar nasal spectrum in a similar way, though using Linear Prediction and without taking vowel context into account. Saito and Itakura (1984) recorded nine male speakers over three years at intervals of six months, saying a single token of the two words /namae/ and /kogeN/ (/N/ here being realised as the velar nasal stop [ng]) on each occasion. Formant frequencies were extracted from Linear Prediction spectra of the three nasal stops, and subjected to an Analysis of Variance. The temporal factor appeared not to be significant, while the speaker factor was highly significant for the higher formants of the bilabial and alveolar nasals, and for all four formants of the velar nasal. They concluded that "the frequency spectra of the nasal consonants are unstable and susceptible to change relative to those of the vowels", but that speaker-characterising information was still available, particularly in the case of the velar nasal (1984: 111). These findings are corroborated by the present experiments.

In summary, there appears to be a good chance that the features of the velar nasal will perform reasonably well in Automatic Speaker Verification, but the unequal dimensionality of the vectors brings considerable potential difficulties. These, and the performance of the velar nasal in speaker verification, are explored in Chapter Seven.

CHAPTER SEVEN

AUTOMATIC SPEAKER VERIFICATION USING NASAL SPECTRAL FEATURES

CHAPTER SEVEN

AUTOMATIC SPEAKER VERIFICATION USING NASAL SPECTRAL FEATURES

7.1. Introduction

In this chapter, the potential of the nasal pole-zero spectrum for automatic speaker verification is explored directly using experimental procedures established in earlier verification studies (e.g. Das and Mohn 1971, Rosenberg and Sambur 1975). These constitute the training of *reference* materials, and the simulation of verification *bids* (see Chapter Two, section 2.3.2). Such simulations are slightly unrealistic, in that the bid materials are pre-recorded under favourable conditions (Chapter Six, section 6.2.3), and not in response to requests from a verification system, nor with the feedback such a system normally provides. The other way in which these trials are unrealistic is that impostor bids are only made using speech from other speakers in the database, with no actual attempts at imitation: in fact, the same utterances may be used both for genuine bids (against a speaker's own reference) and for impostor bids (against the references of others). They are, however, an accepted and convenient way of exploring the usefulness of a set of speech features without the necessity for extensive real-life trials.

The experiments reported here can be divided into two parts: exploration of the *feature set* used for verification (sections 7.3 and 7.4), and experiments in the choice of a *classifier* or comparison procedure (sections 7.5, 7.6 and 7.7; see Chapter Two, section 2.3.2). The feature sets used comprise the peaks and dips of the all-pole and all-zero spectrum, as studied in Chapter Six using analysis of variance, and the raw pole-zero spectrum, which has not been considered so far in this thesis.

The selection of velar nasal stops for these experiments was made in Chapter Three on theoretical grounds, and the performance of stops made at other places of articulation is not explored in this chapter.

7.2. Methods of evaluation

7.2.1. Measures of distance and correlation

Comparisons between reference and test tokens are frequently made on the basis of some measure of *distance* between them. This is readily understood if the feature vector can be conceived of as a point in multi-dimensional space (a *feature space*, defined by the different parameters used).* Thus a single-dimensional vector $\mathbf{x} = [x_1]$, where x_1 is a particular value, for example on the scale of fundamental frequency, marks a point on a straight line. A two-dimensional vector $\mathbf{x} = [x_1,x_2]$ defines a point along each of two axes, x_1 and x_2 (e.g. fundamental frequency and gain). Each such feature vector defines a different point in the space delimited by these two axes. A set of feature vec-

* A geometric interpretation is already implied by the use of the term *vector* to refer to a collection of

tors from a single speaker's utterances should cluster in this space (as they would if plotted on a "scattergram" of gain versus fundamental frequency, for example). Other speakers' feature vectors should show similar clusters, and if verification is to be possible using these two parameters or dimensions, the clusters for different speakers should not overlap but should be separated in this space, with a greater distance (however defined) between the vectors of different speakers than between the vectors of the one speaker. A hypothetical example is shown in Figure 7.1. Each speaker's distribution in the feature space can be represented by the *centroid* of their cluster - that is, the average location of all their feature vectors, typically defined as the mean vector.

The relationship between any two points in this space can be expressed as a *distance*. The simplest distance measure is the length of a straight line (in the 2-dimensional example) between the two points. This is known as the *Euclidean* distance. In the multi-dimensional case (which includes the case of just two dimensions), the Euclidean distance between two points (vectors) $\mathbf{x} = [x_1,x_2,..,x_n]$ and $\mathbf{y}=[y_1,y_2,..,y_n]$ (where n is the number of dimensions or parameters) is given by the formula:

$$d = \left[\sum_{i=1}^{n}|x_i - y_i|^2\right]^{1/2} \qquad (7.1)$$

(Hecker 1971, p.76).

In the process of feature comparison, then, a test vector is compared against the centroid vector of the claimed speaker, and the resulting distance is compared against a *threshold* distance determined for that speaker. This

parameter measurements.

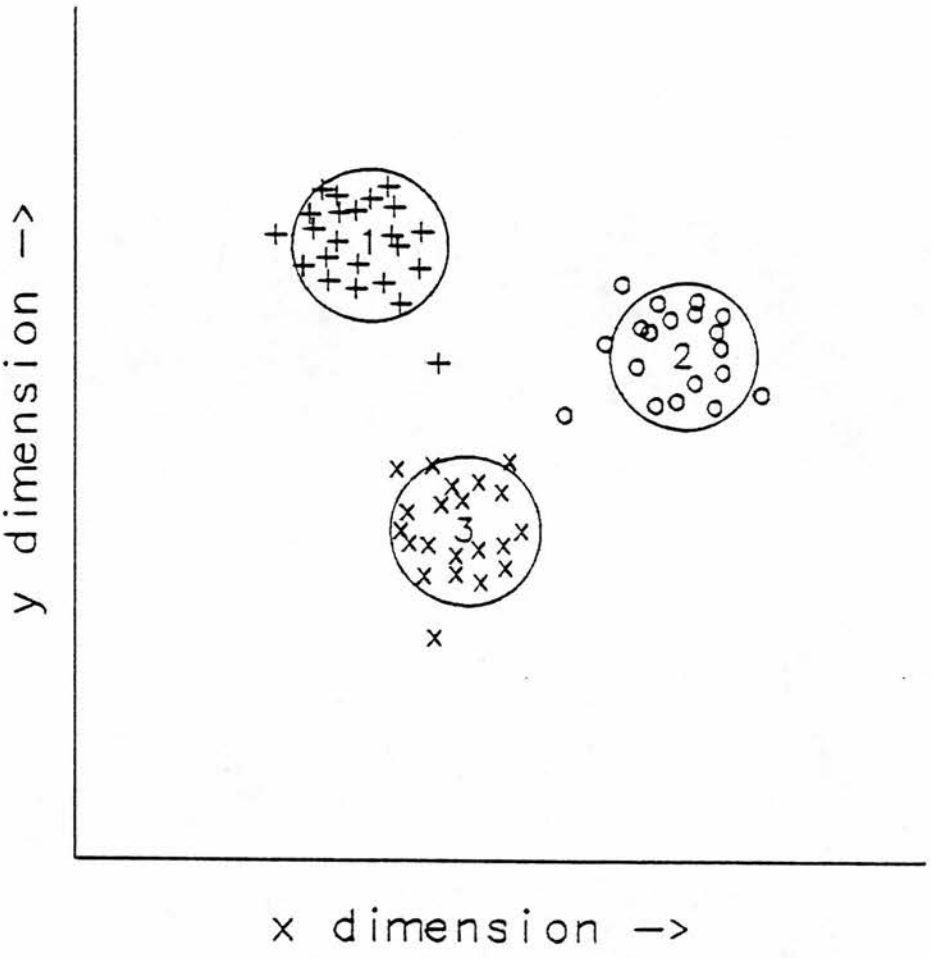

Figure 7.1 Schematic diagram illustrating the clustering of feature vectors in two dimensions. The centroid of each distribution is marked with a numeral.

threshold is chosen so that it "encloses" most (if not all) of the vectors making up that speaker's cluster - in geometric terms, it forms a circle (in two dimensions), a sphere (in three) or a hypersphere (in four or more) around the centroid of the distribution (see Figure 7.1). A test vector falling outside this threshold will be rejected as a valid token from the claimed speaker.

An alternative approach to the use of distance measures is to measure the degree of *correlation* or similarity between two vectors. One such measure

which has been used is the cosine of the angle separating the vectors in multidimensional space (Glenn and Kleiner 1968, Nolan 1983), defined as

$$Z = \frac{X_i \cdot X_j}{(|X_i|\ |X_j|)} \tag{7.2}$$

where X_i and X_j are two different feature vectors. A large value of Z indicates a small angular separation, and is equivalent to a high correlation. An acceptance threshold is set on this value, as with the measures of distance, except that correlations must *exceed* this threshold for speakers to be accepted. Alternatively, the correlation can be expressed in the form $1 - Z$, to maintain the notion of "distance" rather than "similarity".

7.2.2. Measuring the performance of the system

Various methods have been proposed for measuring the performance of speaker verification systems, and comparing the efficiency of different parameter sets. The F-ratio, as applied in Chapter Six to the peak and dip features, can be used to assess the performance of individual parameters (e.g. Wolf 1972, Mohn 1971, Goldstein 1976), but does not give any indication of the error performance of a system using those parameters. A better alternative is the *divergence* (Kullback 1959, Marrill and Green 1963, Atal 1976), which gives a single number indicating the separability of the centroids of all the speakers in a group. One advantage of these measures is that they are independent of the classifier used to give the verification decision, and relate purely to the usefulness of the features themselves.

However, the criterion most directly related to the application, and one which takes classifier performance into account, is the empirically determined *error rate*. The most widely used measure, despite some criticism by Furui (1981a), is the *a-posteriori* Equal Error Rate (see Chapter Two): the point at which the proportion of False Rejections of genuine bids equals the proportion of False Acceptance of impostor bids. This point is generally determined by calculating the distances from each speaker's reference vector to a) their own speech vectors (intra-speaker distances) and b) other speakers' vectors (inter-speaker distances); the resulting distance distributions are used to calculate the error rates (False Rejection and False Acceptance respectively) given different distance thresholds (Figure 2.3, Chapter Two).

The method used to derive the Equal Error Rate is important. Some studies use the training vectors from which the references were formed to provide the "bids" too, but this generally gives an extremely optimistic picture of performance. Foley (1972) has shown that the number of tokens per speaker must exceed the number of elements in each feature vector by a factor of at least three times, if this optimistic bias is to be avoided.

In this study, the dimensionality of the feature vectors reaches a maximum of 128 (the spectral vectors used in 7.3), while the absolute maximum number of vectors available per speaker is 141 (1 session with 15 tokens and 7 with 18). Splitting the data set into separate parts for training and evaluation is therefore recommended. In most experiments in this Chapter this was done by using the first four sessions for each speaker for the reference set, and the

second four as test data. Since one of the female speakers had recorded only four sessions, this speaker was excluded from most analyses.

While it is often recommended that results should be cross-validated by repeating the experiments using the reference data as a test set and forming the references from the test data (Klecka 1980, Das and Mohn 1971, Markel and Davis 1979), it was seen as inappropriate in this case, since it would be unrealistic to form references from speech recorded *after* the test data.

7.3. Peak features versus whole-spectrum parameters

7.3.1. Introduction

It was observed in Chapter Two that spectral information can be used in Speaker Verification in two ways: by treating the spectrum obtained from a token as a feature vector in its own right, and by extracting from the spectrum features such as peak and dip frequencies, bandwidths and amplitudes.

There are advantages in using resonance features such as peak and dip frequencies, as derived in Chapter Six for the nasal pole and zero spectra. The number of elements in the resulting feature vectors is relatively low (a maximum of ten peaks and seven dips were detected in the data presented in Chapter Six), requiring less storage and computation, and allowing more stable references and more robust results from a fixed number of tokens. According to Hunt (1983), peak features also show less susceptibility to distortions introduced by transmission and storage than do the spectral parameters from which they are derived. However, they require an extra processing step, which may

be prone to error, and by their very nature they omit some of the information available for speaker discrimination in the spectrum itself. Conversely, the raw spectral parameters preserve all the available information and are relatively easily derived, but the dimensionality of the feature vectors is high, and there is a great deal of redundancy.

This section begins the exploration of the performance of the nasal pole-zero parameters, therefore, by comparing the two alternative representations.

7.3.2. Whole-spectrum parameters: the pole-zero frequency response

Whereas the peak and dip features studied in Chapter Six were derived from the separate all-pole and all-zero frequency responses respectively, because this enhanced the separability of closely-spaced features, in these experiments the combined pole-zero response (the sum of the two halves) was used. Thus each token was represented by a single 128-element spectral amplitude vector.

Method

A single reference vector was formed for each speaker by calculating the mean vector over all the reference sessions (session 1 to 4) – a total of 69 vectors per speaker. Vowel context was thus ignored (see section 7.4). Intra-speaker and inter-speaker bids were made using the remaining four sessions per speaker, giving a total of 15 * 4 * 18 = 1080 intra-speaker bids and 15 * 14 * 4 * 18 = 15120 inter-speaker bids for the fifteen males, 14 * 4 * 18 = 1008 intra-speaker bids and 14 * 13 * 4 * 18 = 13104 inter-speaker bids for the four-

teen females.

Two classifiers – the unweighted Euclidean distance metric (Equation 7.1) and the correlation measure (Equation 7.2) – were compared. The correlation was expressed in the form $1-Z$, as suggested in section 7.2. The spectral vectors had already been normalized for amplitude differences during the pole-zero analysis; the Euclidean distance could therefore be applied without further normalization.

Results

Table 7.1 shows the calculated Equal Error Rates for the male and female speakers, with the two alternative classifiers. For both sexes, the correlation measure gives marginally better results. Nolan (1983) found very little difference between the two measures, as these results appear to confirm.

7.3.3. Resonance parameters: peak and dip features

The use of the peak and dip features studied in Chapter Six presented two problems. The first is that the features adopted – the peaks of the all-pole spectrum and the dips of the all-zero spectrum – constitute *two* independent

Speaker group	Classifier	
	Euclidean	Correlation
Males (15)	25.053	24.692
Females (14)	24.283	24.119

Table 7.1 Equal Error Rates (%) using 128-dimensional pole-zero spectral vectors

feature vectors, which must somehow be combined. The second problem is that their dimensionality varies from token to token; this prevents the use of both the simple Euclidean distance and measures of correlation, which are only applicable to vectors of equal dimension.

The second problem was overcome somewhat in Chapter Six by the *warping* of each speaker's data to a *prototype* for that speaker, so that all feature vectors had the same number of elements, and so that the elements of all vectors were more or less correctly aligned. In a speaker verification experiment, however, such manipulation is not possible except during the formation of references, since it presupposes that the identity of the speaker of any utterance is known in advance, which is true only of the training phase. Thus the data presented in Chapter Six cannot be used in their present form.

The method of spectral warping presented there does offer a way round the problem of unequal dimensionality, however, which may be particularly applicable to speaker verification and identification. The process of comparison used to align the vectors produces a measure of distance – the sum of the squared differences between the elements of the two vectors over the best warping path (Table 6.8, Chapter Six). This distance thus takes into account not only the size of the differences between corresponding elements, but the extent of the mismatch between the two vectors in terms of the *numbers* of peaks in each, since the distance components from mismatched elements are included in the total. It should therefore give some protection against obtaining a small distance from two vectors having disparate numbers of peaks, even though some of

these peaks were to coincide exactly.

The problem of combining the feature vectors remains. One disadvantage of this measure is that it prevents the two output vectors from the pole-zero analysis from being combined into a single vector, since the warping must only be allowed to compare like with like. It is possible, however, to combine the *distances* produced. Several methods of combining independently derived distance measures have been suggested in the literature. One method is to use each distance as the basis for a verification decision, and to combine these decisions in a logical fashion (Mohankrishnan et al. 1982). Alternatively, the two distances could be normalised in some way and combined to give a single distance on which the verification decision is based. A useful technique of this second type, which has the advantage of preserving more information about each bid, is suggested by Sutherland (1989): each distance is related to the distributions of the intra-speaker and inter-speaker distances derived from the training data for that parameter set, to give a measure of the confidence with which the test token can be verified as a token from the claimed speaker. This measure has been termed a *certainty score*. The intra-speaker distances produced during training are converted to a cumulative density function, which is then scaled and inverted to indicate the (falling) probability of false rejection with an increasing distance threshold, $P(FR)$ (Figure 7.2); the inter-speaker distances are also converted to a cumulative density function, to represent the (rising) probability of false acceptance with an increasing threshold, $P(FA)$. The two curves are related into a single function by finding the difference

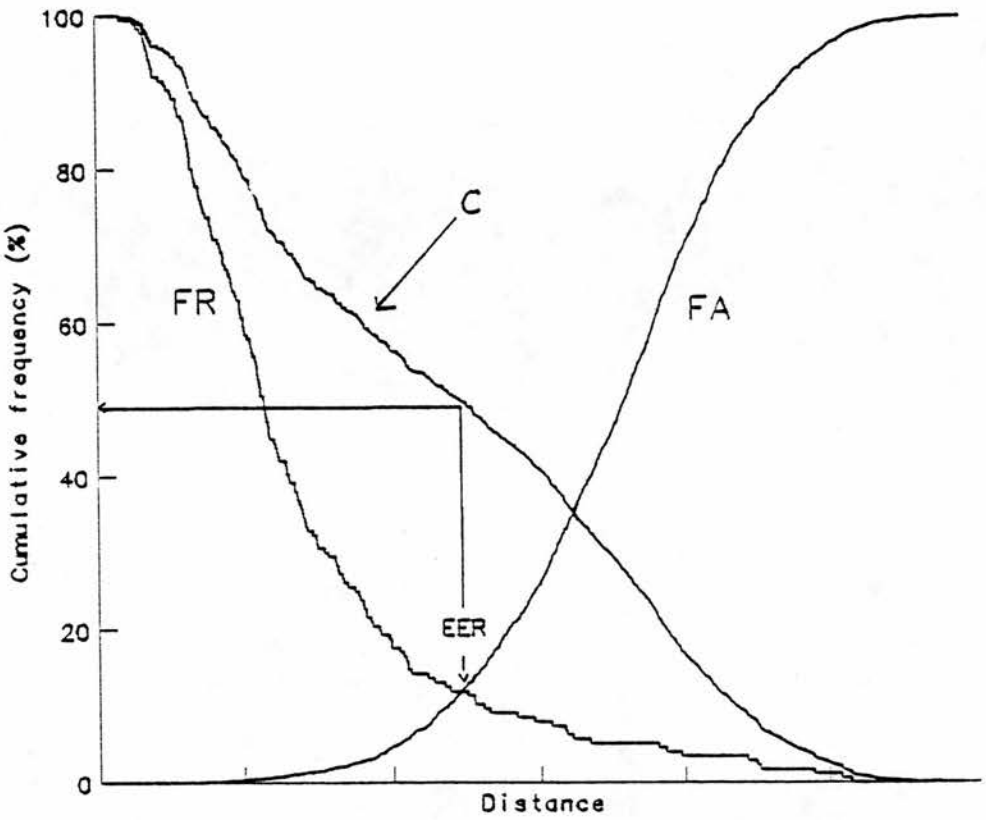

Figure 7.2 Derivation of the certainty score from cumulative intra-speaker (FR) and inter-speaker (FA) distance distributions. The EER point corresponds to 50 % certainty.

between them at each point on the distance axis, and scaling these differences to range between 100 % (total certainty of genuine identity) for distances below the lowest inter-speaker distance observed, and 0 % (total certainty of a false claim) for distances above the highest intra-speaker distance observed, as in Equation 7.3:

$$c = \frac{P(FR) - P(FA) + 100}{2} \tag{7.3}$$

Distances obtained during a bid can then be related to the corresponding certainty function, to give a measure of confidence independent of the units of distance involved; these certainty scores can then be combined by averaging (with or without additional weighting dependent on the efficiency of the separate parameter sets), to give a final certainty score which is compared against a combined threshold (50 % being the threshold corresponding to the Equal Error Rate).

In this section, therefore, the performance of the peak and dip parameter sets was studied using spectral peak warping on each set separately, before the results were combined using the certainty scores of each set.

Method

A single pair of references was formed for each speaker by finding the prototypical peak and dip vectors from their reference data (sessions 1 to 4) – a total of 69 vectors per speaker. Vowel context was again ignored. Because of the reduced amount of data available for each speaker, *no* constraint was placed on the number of elements in the resulting prototypes – that is, all 69 reference vectors were candidates for the prototype; otherwise, the same method was used to select the prototype as was set out in Chapter Six, that of minimizing the average variance over the resulting warped vectors. Intra-speaker and inter-speaker bids were made using the remaining four sessions per speaker, but some tokens produced no dip frequency vector, and were thus excluded from the study of the dip parameters (though their corresponding peak vectors were used). The totals were 1078 peak/920 dip intra-speaker bids and

15120 peak/12880 dip inter-speaker bids for the fifteen males; and 1008 peak and 915 dip intra-speaker bids and 13104 peak and 11895 dip inter-speaker bids for the fourteen females. The Euclidean distance produced by warping the test token to the reference formed the basis of the decision.

The combined performance of the two parameter sets was studied by converting each distance to a certainty score using the appropriate certainty function, and averaging the two certainty scores obtained for each token. For this part of the experiment, those tokens which gave no dip vector were excluded completely, since both distances were required for the combination. The total numbers of bids were therefore 920 intra-speaker and 12880 inter-speaker for the males, and 915 intra-speaker and 11895 inter-speaker for the females.

Results

The two feature sets were examined independently, and the equal error rate for each was calculated. These rates, together with the EER obtained on the combined parameter set, are given in Table 7.2. It is clear that the performance of the peak and dip features is much poorer than that of the unreduced pole-zero spectrum, being little better than chance. In the case of the females,

Speaker group	Parameter set		
	Peaks	Dips	Combined
Males (15)	50.396	44.251	46.068
Females (14)	49.945	48.466	46.557

Table 7.2 Equal Error Rates (%) using warped peak and dip frequency vectors

combining the parameter sets produced a small improvement in performance, but the results remained poor.

7.3.4. Discussion

The performance of the pole-zero spectral parameters, at around 25 % EER or less, is encouraging, and compares well with Glenn and Kleiner's (1968) *identification* error rate of 57 % using the filter-bank spectrum of alveolar nasals in a group of mixed sex; they reduced this rate by forming bids by averaging several tokens and dividing the population into smaller groups. However, the poor performance of the peak and dip features is difficult to reconcile with the promise of their F-ratios in Chapter Six. One possibility is that the full 128-element spectral vectors have an advantage over the smaller peak-dip feature sets simply because of their higher dimensionality. However, later studies in sections 7.6 and 7.7 will show that similar results can be obtained with a much smaller number of parameters, using Canonical Analysis. Another is that the process of peak and dip measurement has discarded too much information, such as bandwidths and amplitudes. It is also possible that the method of distance measurement, using peak profile warping (6.5.1), has proved inappropriate: because speakers' tokens vary so much in dimensionality, many high intra-speaker distances may result from the mismatch in size, rather than from actual differences in the frequency values of corresponding peaks, which were shown to be relatively stable in the F-ratio analysis of Chapter Six. It is possible that a more constrained version of this distance measure might perform better. Alternatively, a representation of spectral

peaks which guarantees the same number of elements in every vector might be used; one possibility is the use of *generalized centroids* (Crowe and Jack 1987), where the spectral vector is divided into a number of sections of varying length, each represented by its centroid (mean frequency value); the lengths of the sections are determined by minimising the sum of the resulting squared error terms within each section. For application to the pole-zero parameters, however, the algorithm would have to be modified to cope with the location of *dips* in the all-zero spectrum.

In view of the results of this experiment, then, the whole pole-zero spectral vector is used in preference to the peak and dip features in all subsequent experiments.

7.4. Effects of vowel context on the use of the pole-zero spectrum

7.4.1. Introduction

The context of the nasal stop, though ignored in the preceding section, was shown in Chapter Six to have a significant effect on the values of peak and dip frequencies, and it is therefore likely that it should be taken into account in the formation of references and the comparison of bid tokens: including tokens from different contexts may increase the spread of a speaker's distribution in the features space, requiring the use of higher thresholds than would otherwise be necessary. Its effect on the verification error rate given by the whole pole-

zero spectrum is therefore studied here.

Method

Exactly the same experimental paradigm was followed as in section 7.3.2, except that each speaker was provided with *three* references, each being the mean of the 23 training vectors spoken in a given vowel context (*ing*, *ang*, *ung*). Vectors from the last four sessions again formed the test data, but each test bid was made only against a reference from the same vowel context, giving the same number of bids overall. The assumption was made that speakers in a real-life system would cooperate in saying the vowels required of them. Both the Euclidean distance and the correlation classifiers were used.

Results

The Equal Error Rate was calculated for each vowel context separately from the vowel-dependent intra-speaker and inter-speaker distance distributions, and for the system as a whole from the pooled distributions. The resulting values are given in Table 7.3. The pooled rates should be compared with those given in Table 7.1.

Discussion

It is clear that the use of vowel-dependent references and bids has reduced the overall Equal Error Rate in all cases. The advantage of the correlation measure over the Euclidean distance is maintained. The vowel-dependent error rates show small differences, but the fact that the ordering of the three vowel contexts for the male speakers ($ang < ing < ung$) is the reverse of that for

Speaker group	Context	Classifier	
		Euclidean	Correlation
Males (15)	ing	21.996	21.727
	ang	20.000	19.970
	ung	22.016	21.787
	pooled	21.339	21.144
Females (14)	ing	20.822	20.524
	ang	21.021	21.021
	ung	19.369	19.127
	pooled	20.517	20.437

Table 7.3 Equal Error Rates (%) by vowel context, using 128-dimensional pole-zero spectra, vowel-specific references

the females (*ung* < *ing* < *ang*) suggests that these differences are not significant, and that no one vowel context performs any better than the others.

While the overall results are better, the disadvantage of taking vowel context into account is that storage and computation requirements are increased for a working system, since three references must be stored for each speaker, along with three distance thresholds. It is also possible that more training data would be required to produce stable references, though the reduction in intra-speaker variance obtained by restricting the vowel context for each reference probably compensates for this.

These results confirm that vowel context does influence the spectrum of the velar nasal, presumably through coarticulatory variation in tongue and larynx position. While it might have been expected that the low vowel context would give less reliable results, since there is more chance of speakers

producing an incomplete velar closure and of introducing coupling to the oral cavity, there is no evidence of this.

7.5. Improvements to the design of a classifier

7.5.1. Introduction

This section examines how the design of the classifier – in particular, the application of appropriate statistical tools – can influence the results obtained using the pole-zero spectrum. Much of the progress made in speaker recognition in recent years has been in the understanding and application of statistical techniques for improving classification performance, and if the pole-zero spectrum – or any other feature vector – is to be judged fairly, it is important that some of the improvements offered by the use of such techniques are adopted.

7.5.2. The use of variance-based weightings in the Euclidean distance classifier

The standard Euclidean distance is a measure of the straight line distance between two points in space, as observed in section 7.2. The use of this distance gives equal weight to all dimensions along which two points differ. It may be the case, however, that some dimensions have greater variation in their observed values than others. Thus the distance components measured along these dimensions will be higher automatically, and will contribute disproportionately to the total distance between the points.

Figure 7.3 shows an example in two dimensions or variables. These variables have unequal variances, the x values showing a much greater spread than the y values for both speakers. Thus a large distance on the x dimension is less significant than a large distance on the y dimension. The imposition of a circular threshold, such as that imposed on the distributions shown in Figure 7.1, will result in many genuine bids being rejected, and many "impostor" bids from other speakers being accepted. To make the use of a circular threshold

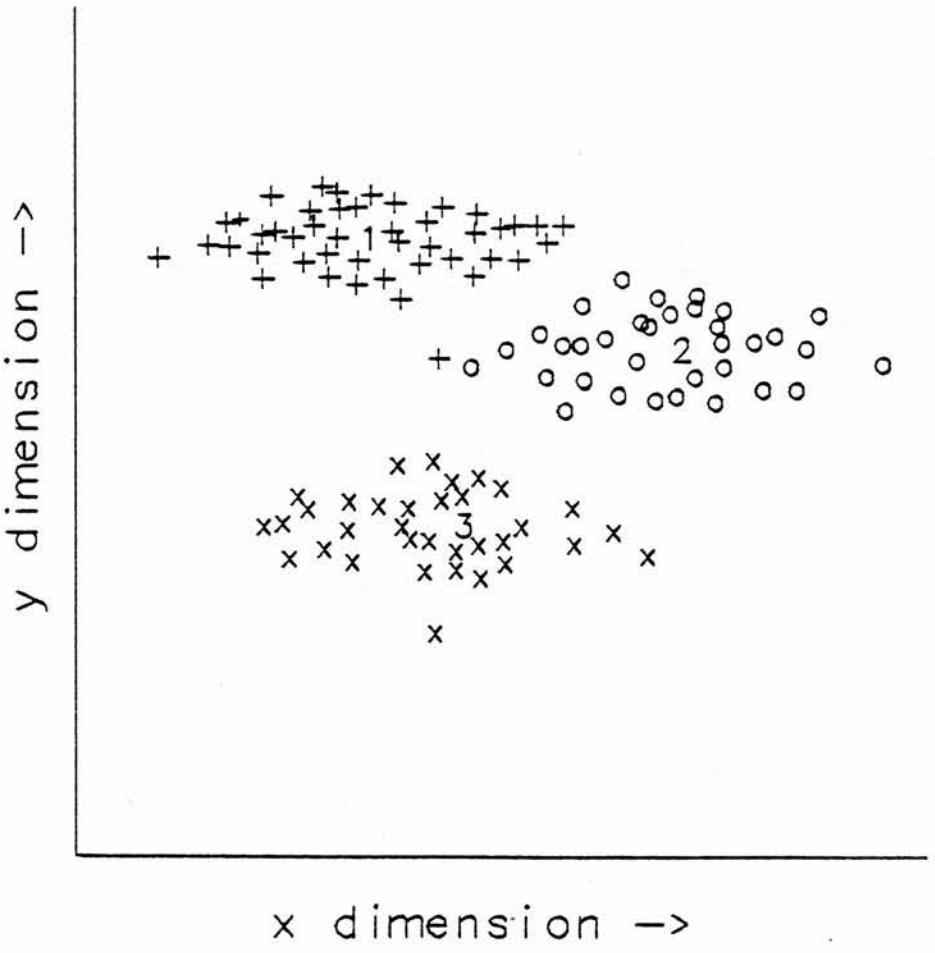

Figure 7.3 Schematic illustration of the clustering of feature vectors in two dimensions with unequal variances

possible, then, the distributions must be made approximately circular again, by taking into account the variances of each parameter. This is done by weighting each component of the Euclidean distance by the inverse of the variance observed along that dimension. Equation 7.1 therefore becomes

$$d = \left[\sum_{i=1}^{n} |x_i - y_i|^2 \cdot v_i^{-1}\right]^{1/2} \qquad (7.4)$$

where v_i is the variance of parameter i. The variances can be estimated over the training vectors of each speaker. Speaker-specific variances may be used, as a way of coping with differences between speakers. Alternatively, the population variance (the pooled intra-speaker variance) can be estimated.

Method

The weighted Euclidean distance given in Equation 7.4 was applied to the existing pole-zero spectrum references and bids, as in section 7.4. Both general and vowel-specific references were formed. Speaker-specific variance weightings were calculated over the 69 training vectors for each speaker, or in the case of the vowel-specific references, over the 13 training vectors from a given vowel context. For the vowel-specific trials, the distance distributions from each vowel were pooled before Equal Error Rates were calculated.

Results

Table 7.4 gives the results for the general and the (pooled) vowel-specific trials. These represent a small improvement on those obtained using the unweighted Euclidean distance in the case of the males, but the females show a very small deterioration. These figures are generally not quite as good as those

obtained using the correlation classifier, however, which already takes differences in variance into account and does not require a separate weighting vector.

7.5.3. The application of Canonical (Linear Discriminant) Analysis

The use of variance-based weightings is helpful, but it still makes two assumptions which have major consequences for the performance of a system: that the 128 parameters make contributions of equal value to the discrimination between speakers, and that they are *uncorrelated.* The use of parameters which do not contribute significantly to discrimination is at best wasteful of storage space and computing time, but at worst it can be detrimental to the performance of a classifier, particularly where the number of dimensions is large and the number of training tokens is small. It is therefore wise to remove from the classification any parameters which do not contribute, or whose contribution is negligible. A high degree of correlation among parameters is also to be avoided, even where individually they contribute significantly to the discrimination: this not only constitutes more redundancy, but can also be detrimental to performance because the shape of speakers' distributions in

Speaker group	Reference type	
	Global	Context-dependent
Males (15)	22.171	19.272
Females (14)	24.359	21.852

Table 7.4 Equal Error Rates (%) for 128-dimensional pole-zero spectra using weighting based on individual speaker variances

the feature space, and therefore the appropriateness of the threshold on distance, is affected. Figure 7.4 shows a hypothetical two-dimensional case in which the two parameters show a high degree of positive correlation; speakers' distributions therefore take the shape of an ellipse, rather than a circle (even when variances are identical), and a circular threshold is therefore inappropriate.

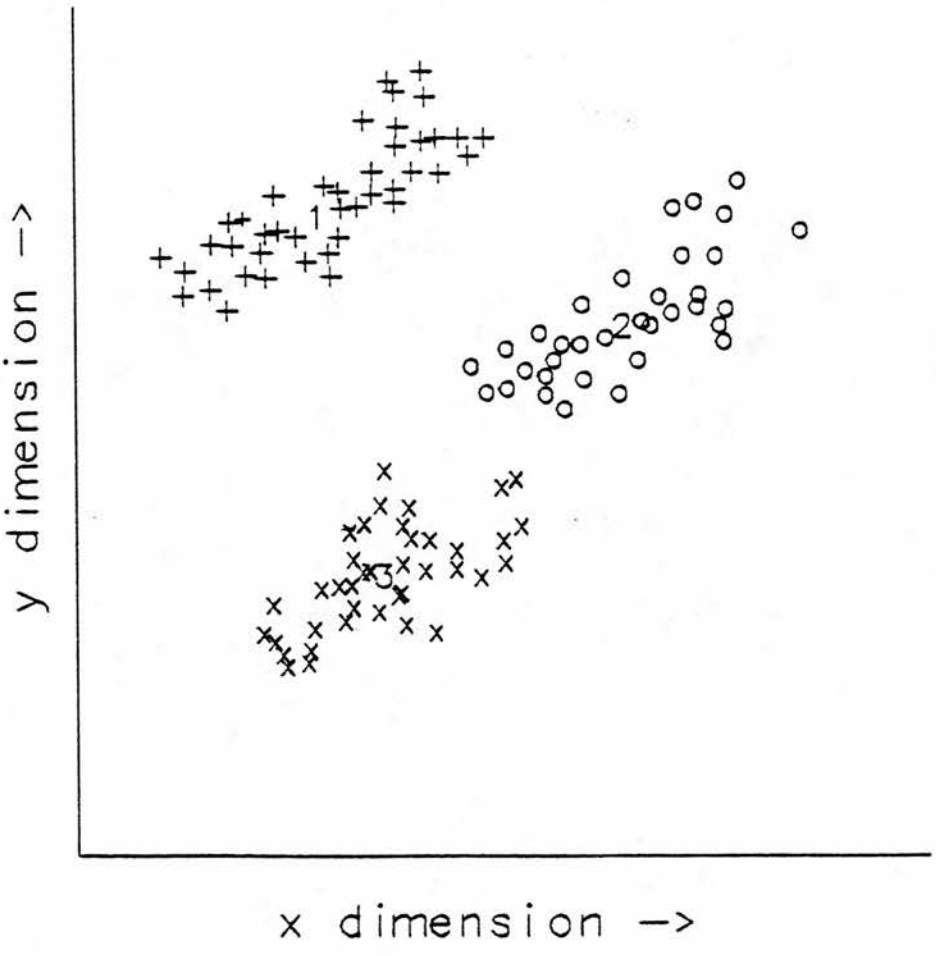

Figure 7.4 Schematic illustration of the clustering of feature vectors in two dimensions showing a high positive correlation

The selection of a subset of features can thus be helpful. The F-ratio has frequently been used to achieve this (e.g. Mohn 1971, Wolf 1972), but it has the drawback that it fails to take correlations into account. It is useful, however, to see how the individual contributions of the parameters compare. An F-ratio analysis of the 128 spectral parameters was therefore carried out on a subset of the database (all vectors from the first two sessions for 15 male and 15 female speakers). The F-ratios are plotted on a frequency-scaled axis in Figure 7.5. It is clear that there are large differences among the parameters in their possible individual contribution to speakers' discriminability. It is also interesting to note that the high points appear to coincide with certain of the peak and dip locations studied in Chapter Six, especially in the region of the 1000 Hz peak.

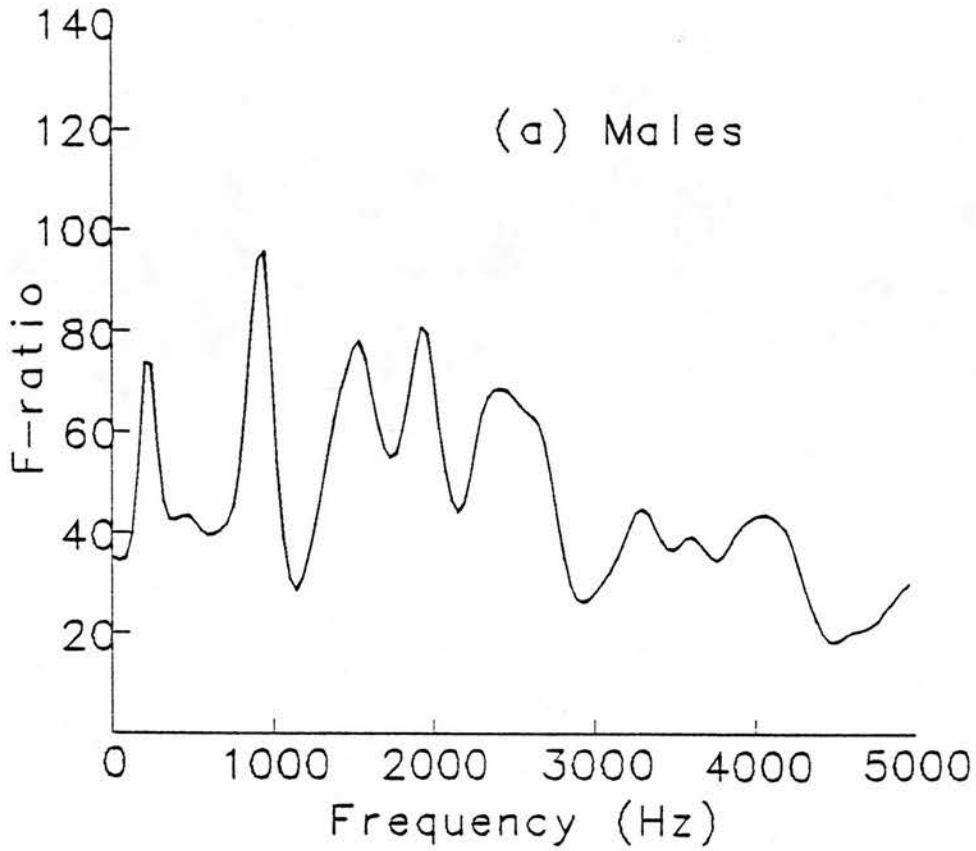

Figure 7.5 F-ratio values by spectral bin frequency for (a) 15 male and (b) 14 female speakers, 128-dimensional pole-zero spectrum; 2 sessions per speaker, vowel contexts pooled.

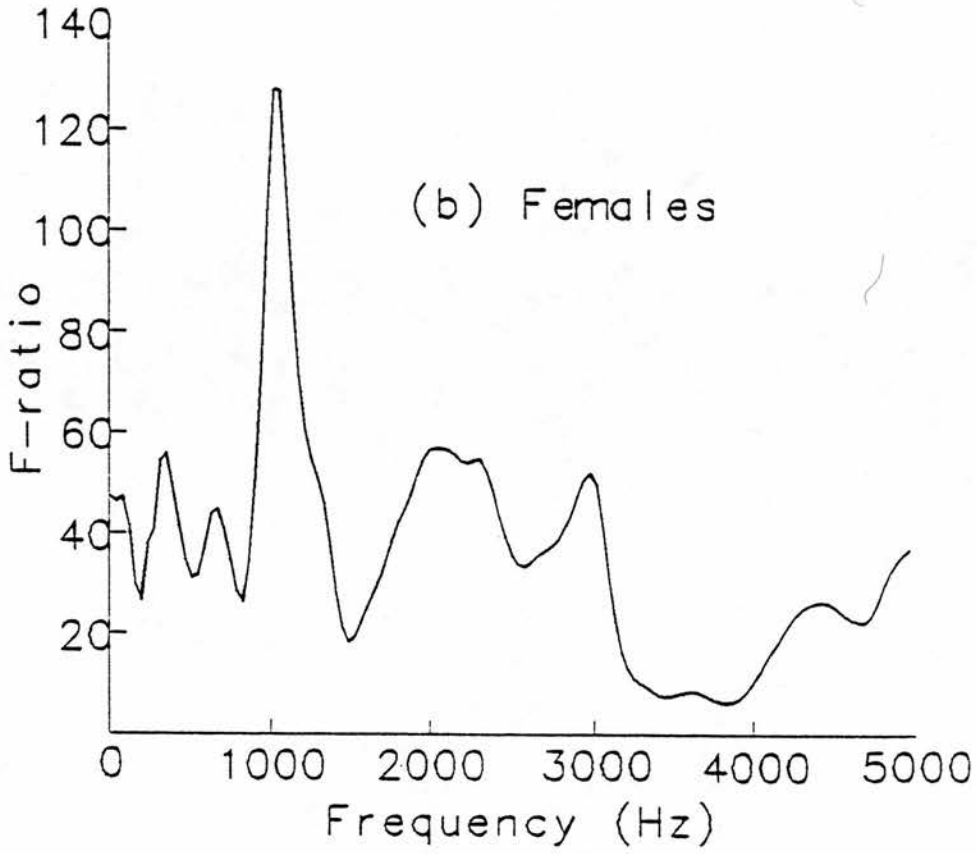

Figure 7.5 F-ratio values by spectral bin frequency for (a) 15 male and (b) 14 female speakers, 128-dimensional pole-zero spectrum; 2 sessions per speaker, vowel contexts pooled.

The degree of correlation among the 128 spectral parameters was assessed on the same subset of data. Product-moment correlations were calculated between each spectral parameter and all other parameters, to give a 128-by-128 dimension correlation matrix. A portion of this matrix is shown in Table 7.5; the diagonal elements represent the correlation of each variable with itself, and are therefore equal to 1. To summarise the extent of inter-parameter correlation, this matrix was reduced by averaging the correlations between parameters one element apart, those between parameters two elements apart and so on. The average correlation coefficients for pairs of parameters one, two , three, four and five elements apart are given in Table 7.6. These show that there is a very high degree of correlation between adjacent elements, and that this correlation declines as the separation between elements increases. Such a result is to be expected, since the smoothed frequency spectrum of speech contains few discontinuities, and amplitude values are generally predictable from the values of lower frequency bins over a short span. The size of these correlations, how-

	Spectral bin number						
	1	2	3	4	5	6	7
1	1.000						
2	0.997	1.000					
3	0.972	0.987	1.000				
4	0.874	0.907	0.963	1.000			
5	0.642	0.695	0.800	0.931	1.000		
6	0.324	0.389	0.526	0.731	0.927	1.000	
7	0.043	0.110	0.256	0.492	0.760	0.937	1.000

Table 7.5 A portion of the inter-correlation matrix for the 128 spectral parameters (male speakers).

Speaker group	Interval between parameters				
	1	2	3	4	5
Males	0.98	0.92	0.84	0.74	0.63
Females	0.98	0.91	0.82	0.71	0.60

Table 7.6 Mean correlation coefficient between elements in pole-zero spectrum with increasing separation; 15 males, 15 females; 2 sessions each

ever, indicates that there is considerable scope for data reduction in this representation.

In this section, the effects of statistical optimisation using Canonical analysis are studied. Canonical analysis (James 1985) is a technique which allows both aims – the removal of redundant parameters and the elimination of inter-parameter correlation – to be achieved very efficiently. A small number of new, uncorrelated parameters is derived from existing parameters by combining them into a weighted sum. The relationship between the original parameters and the new parameters is defined by a set of *canonical discriminant functions*, which can be used either to explore the nature of inter-speaker differences, or as part of a classifier. The maximum number of uncorrelated parameters – that is, the maximum number of functions that can be derived – is equal to one less than the number of groups (speakers). This is because this is the number of dimensions needed to achieve maximum separation of the groups.

The use of canonical analysis has been shown to be beneficial in other studies (e.g. Cheung 1978), but its use has also been criticised (Mohn 1971) as

being efficient only for the original data used to design the classifier.

Method

The aim of this experiment was to study the effect on error rates of using an optimal set of uncorrelated variables in place of the existing 128 spectral parameters. This set of variables was derived from the spectral data using the BMDP Stepwise Discriminant Analysis (P7M) program (Dixon 1985). A set of canonical discriminant functions was derived separately for each sex group, from a small amount of training data comprising the first two sessions from each speaker (15 males and 14 females). To ensure a stable analysis with only two sessions of data per speaker, tokens from the three vowel contexts were pooled in the derivation of the discriminant functions.

Two options for the derivation of the functions were examined: in the first, every one of the 128 parameters was used in the combination, no matter how small its contribution, while in the second the program was allowed to select the optimal subset of parameters for inclusion in the functions using a "stepwise" procedure (Klecka 1980) in which variables are only entered if their entry improves discrimination by a significant amount, and can be removed if the entry of other variables makes them redundant. It should be noted that the number of discriminant functions derived – and therefore the number of output variables – is the same (one less than the number of speakers) in each case: it is the number of original variables used in these functions that differs. For the males, the stepwise procedure selected 26 of the original 128 variables, while for the females the figure was just 25. It was found that the error rates

produced by the two procedures (full and stepwise) were virtually identical, and only the stepwise functions are presented here.

Each discriminant function takes the form of a set of weighting coefficients, one for each selected variable, and a constant. Coefficients for the variables to be excluded are set to zero. Any input vector can then be projected onto the new dimensions defined by a set of functions. Each function is applied to the input vector in turn, and produces a single output value, the score of that input vector on the corresponding dimension. The output vector thus has as many dimensions as the number of functions defining the new feature space. This operation is illustrated in Figure 7.6.

The resulting 14 functions for the males and 13 functions for the females were used to project the general *and* the vowel-specific references derived in section 7.3 on to the new feature space as above, giving new references of 14 and 13 elements respectively. Since the discriminant functions are *linear* combinations of the existing variables, projection of the mean vector in this way is equivalent to the individual projection and averaging of all the training vectors. These references had been formed from the first *four* sessions of data, and thus only part of this reference material had also contributed to the derivation of the discriminant functions. The remaining four sessions' tokens were then projected in a similar way to form intra-speaker and inter-speaker bids, as in section 7.3. The unweighted Euclidean distance classifier and the correlation measure were used: there was no need for variance-based weighting in this

Given a set of P canonical functions composed of j weighting coefficients and a constant K:

$$F_1\colon c_{1,1},\ c_{1,2},\ \cdots\ c_{1,j}\quad K_1$$

$$F_2\colon c_{2,1},\ c_{2,2},\ \cdots\ c_{2,j}\quad K_2$$

...

...

$$F_p\colon c_{p,1}\ c_{p,2}\ \cdots\ c_{p,j}\quad K_p$$

and an original feature vector X of j elements

$$X\colon x_1\ x_2\ \cdots\ x_j$$

the new vector X' is given by:

$$X'_1 = F_1 * X$$

$$= (c_{1,1} * x_1) + (c_{1,2} * x_2) \cdots + (c_{1,j} * x_j) + K_1$$

$$X'_2 = F_2 * X$$

$$= (c_{2,1} * x_1) + (c_{2,2} * x_2) \cdots + (c_{2,j} * x_j) + K_2$$

...

...

$$X'_p = F_p * X$$

$$= (c_{p,1} * x_1) + (c_{p,2} * x_2) \cdots + (c_{p,j} * x_j) + K_p$$

Figure 7.6 Projection of original feature vectors on to the orthogonalized feature space using canonical discriminant functions.

experiment, since the projected variables have equal variances.

Results

The resulting Equal Error Rates are given in Table 7.7. These show that the use of canonical variables has improved performance considerably compared

with that achieved in sections 7.3 and 7.4, particularly in the case of the correlation classifier. The vowel-specific references keep their advantage in this representation, even though a single set of discriminant functions was used to project the three references for each speaker. It is possible that their performance might be improved by the derivation of separate discriminant functions for each vowel context, but this would require extra training data and increase computation considerably.

7.5.4. Effects of a reduction in dimensionality

The canonical analysis used in this section has reduced the dimensionality of the vectors being used from 128 to 14 or less, giving a considerable reduction in the storage required for reference information. It may be desirable, however, to reduce the dimensionality further if possible, by selecting the best subset of features from the new feature set – the set of 14 (13) canonical discriminant functions. Canonical analysis makes such a selection very easy, since the functions are already ordered by their discriminatory power: the first function to be

Speaker group	Reference type	Classifier	
		Euclidean	Correlation
Males (15)	Global	22.187	13.908
	Context-dep	21.018	12.815
Females (14)	Global	18.882	14.833
	Context-dep	17.626	13.845

Table 7.7 Equal Error Rates (%) for 14-dimensional (male) and 13-dimensional (female) orthogonalized spectral vectors derived by Canonical Analysis

derived is the one which maximises the separation between the speakers, while all subsequent functions contain less and less speaker-discriminating information. It is possible to gauge the discriminatory power of each function by comparing the *eigenvalues* of the functions after they have been normalized to sum to 100 per cent (Klecka 1980). Table 7.8 shows the relative percentage contribution to the discrimination of each function for the male and female discriminant function sets, calculated from their eigenvalue. Much of the available discrimination is concentrated in the first few functions, the later discriminant functions carrying only a very small proportion of the total. Table 7.8 also shows the *canonical correlation* of each function – that is, its correlation with a dummy variable denoting group membership or speaker identity (James 1985).

Function	Males		Females	
	Rel. Perc.	Can. Corr.	Rel. Perc.	Can. Corr.
1	19.754	0.910	27.697	0.945
2	14.689	0.884	17.969	0.919
3	13.430	0.875	14.563	0.903
4	10.162	0.844	12.503	0.890
5	9.486	0.835	7.408	0.832
6	8.198	0.816	6.502	0.815
7	5.907	0.768	4.264	0.751
8	5.259	0.749	2.740	0.674
9	3.898	0.698	2.003	0.615
10	3.231	0.664	1.822	0.597
11	2.180	0.589	1.258	0.526
12	1.925	0.565	1.016	0.485
13	1.170	0.471	0.254	0.268
14	0.711	0.384		

Table 7.8 Relative percentage contribution to discrimination and canonical correlations of the 14 male and 13 female canonical discriminant functions

Correlations are high for the first few functions, but fall off gradually, with the last four functions in each case having correlation coefficients of less than 0.6.

These relative values suggest that quite a high level of performance can be achieved using less than the full set of discriminant functions. A small experiment was therefore carried out, on five male speakers, to assess the effect on the Equal Error Rate of varying the number of dimensions used.

Method

References from the first five male speakers were projected onto the new feature space using the set of 14 discriminant functions, as in the preceding subsection. Data from their remaining four sessions were also projected on to the new feature space. Bids were then made using an increasing number of dimensions of the reference and test vectors: that is, the first set of bids were made using only the first element of the reference and test vectors, with additional elements being added in turn, up to the maximum of 14 elements as used in section 7.5.3. The unweighted Euclidean distance and the correlation classifiers were used, with both general and vowel-specific references.

The Equal Error Rate given *without* canonical analysis was also calculated for this group of speakers using the original 128-element spectral vectors, to allow a valid comparison to be made with the performance of the canonical functions.

Results

Figure 7.7 shows the changes in the Equal Error Rate for both general and vowel-specific references, using both types of classifier. The Equal Error Rates obtained for this group of five speakers using the original 128-element vectors are marked with arrows for comparison. It can be seen that the use of only the first canonical function gives an EER of between 35 % and 40 %, well above

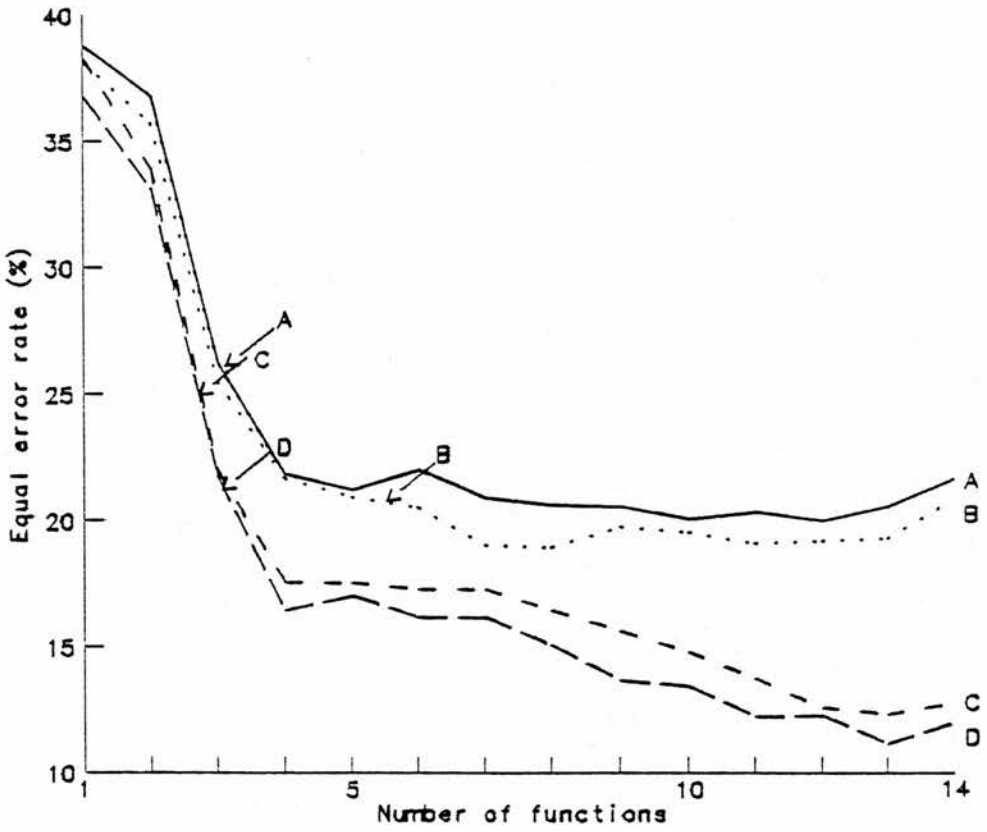

Figure 7.7 Effects on Equal Error Rate of increasing the number of canonical variables used in verification trials; five male speakers only. Curves: (A) Euclidean classifier, pooled contexts; (B) Euclidean classifier, context-dependent; (C) correlation classifier, pooled contexts; (D) correlation classifier, context-dependent.

that produced by the original data. However, only three or four dimensions are needed before the error rates drop below the levels given by the original parameters. After this point, the improvements in performance appear to flatten out in the case of the Euclidean classifier (curves A and B), but the performance of the correlation classifier (curves C and D) continues to improve, albeit more slowly. The reasons for this difference in behaviour are not clear, but it demonstrates that the choice of a classifier can have a substantial effect on the performance of a system.

7.5.5. Discussion

This section has shown that the performance of velar nasal spectra for speaker verification can be improved considerably by the application of appropriate statistical techniques. The greatest benefit has come from the use of Canonical Analysis to remove redundancy and correlation. The cost of this improvement, however, is increased computation and, ideally, additional training data to estimate the discriminant function coefficients. It is also not clear how a system using these functions would cope with the enrolment of a new speaker: it is possible that the discriminant functions would have to be re-evaluated each time.

7.6. Speaker differences in Automatic Speaker Verification performance

7.6.1. Introduction

The following two sections, 7.6 and 7.7, address questions relating to the effects of the speakers themselves, rather than to the parameters and classifier design.

Differences among speakers in their performance on speaker verification systems have been noted in several studies (e.g. Doddington 1985, Rosenberg and Shipley 1983). It has been observed that most false rejections of genuine bids occur in a small percentage of the population. Doddington notes that only one quarter of the population had rejection rates greater than the average (mean), and that the "typical" user – one showing the *median* rejection rate – had a rejection rate of just half of the mean. The distribution of errors of false rejection was therefore examined in the present database.

7.6.2. The distribution of errors across speakers

The number of errors of false rejection in the data presented in section 7.5 was calculated for each speaker separately at the global Equal Error Rate threshold estimated in 7.5.3 (that is, 15 male speakers and 14 female speakers, 14-dimensional orthogonalized parameters, vowel-specific references). To gain an idea of the spread of these errors, the number of false rejections for each speaker was expressed as a percentage of the *total* number of false rejections occurring for each sex group. The speakers were then ranked in order of increasing FR percentage. These percentages are shown in Figure 7.8. It should be noted that these are not False Rejection *rates* (which are expressed relative to the total number of genuine *bids*).

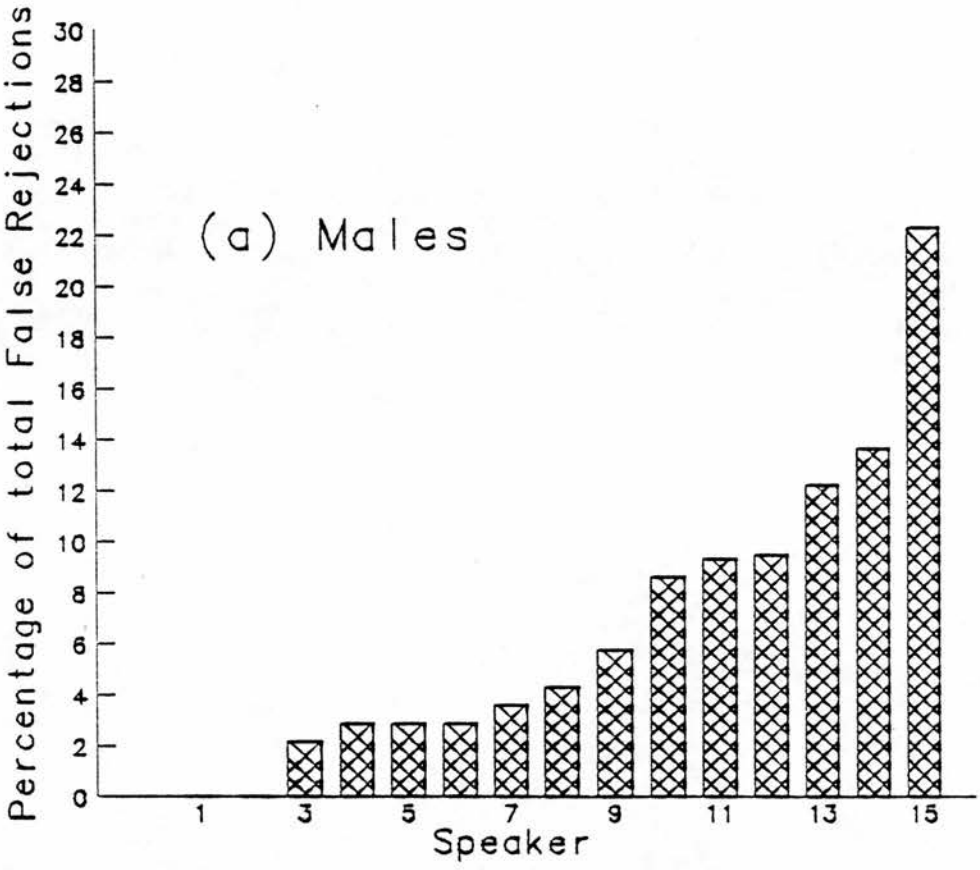

Figure 7.8 Percentage of total False Rejection errors by speaker, for (a) 15 males and (b) 14 females.

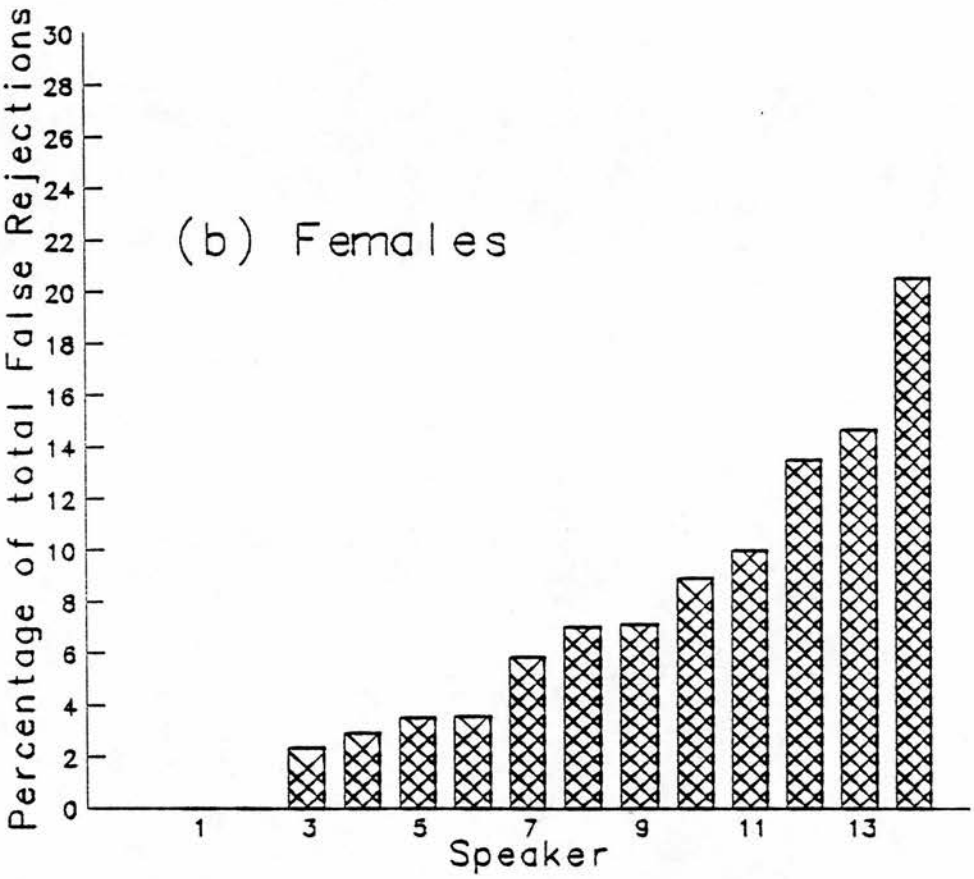

Figure 7.8 Percentage of total False Rejection errors by speaker, for (a) 15 males and (b) 14 females.

It can be seen that there are large differences among speakers in the distribution of errors of false rejection, as noted in other studies. Such differences presumably occur because of differences in the *consistency* of people's speech performance. A high FR rate suggests that a speaker's reference is not representative of the range of tokens they typically produce, or that the global threshold used here, established from the distance distributions of all speakers, is too low for that speaker. In such cases, a *speaker-specific* threshold may be required.

Speaker-specific distance thresholds have been proposed as a way of coping with differences in speakers' variability. Instead of a single global threshold, a separate threshold is estimated for each speaker on the basis of his or her own intra-speaker distances only and the distances given by other speakers' bids against his or her reference. One problem with this is that the amount of data available for calculating these thresholds is necessarily much smaller, and the resulting thresholds may be cruder and less robust than a global estimate. Another, possibly more serious, problem is that the wider threshold given to a speaker whose performance varies may give a greater False Acceptance rate to impostor bids, with the result that the overall performance of the system is unchanged, or even slightly worse.

7.6.3. Speaker-dependent thresholds

To investigate the effect of speaker-specific thresholds on the distribution of errors, individual Equal Error Rate thresholds were estimated for the 15 male and 14 female speakers used above, and the corresponding False Rejection

rates were calculated. To make the comparison easier, these FR rates, and those obtained using the global thresholds (Figure 7.8), were expressed in cumulative form by calculating each speaker's share of the total number of errors, ranking the speakers and plotting the cumulative percentage of errors accounted for as each speaker was included. These cumulative density functions are presented in Figure 7.9.

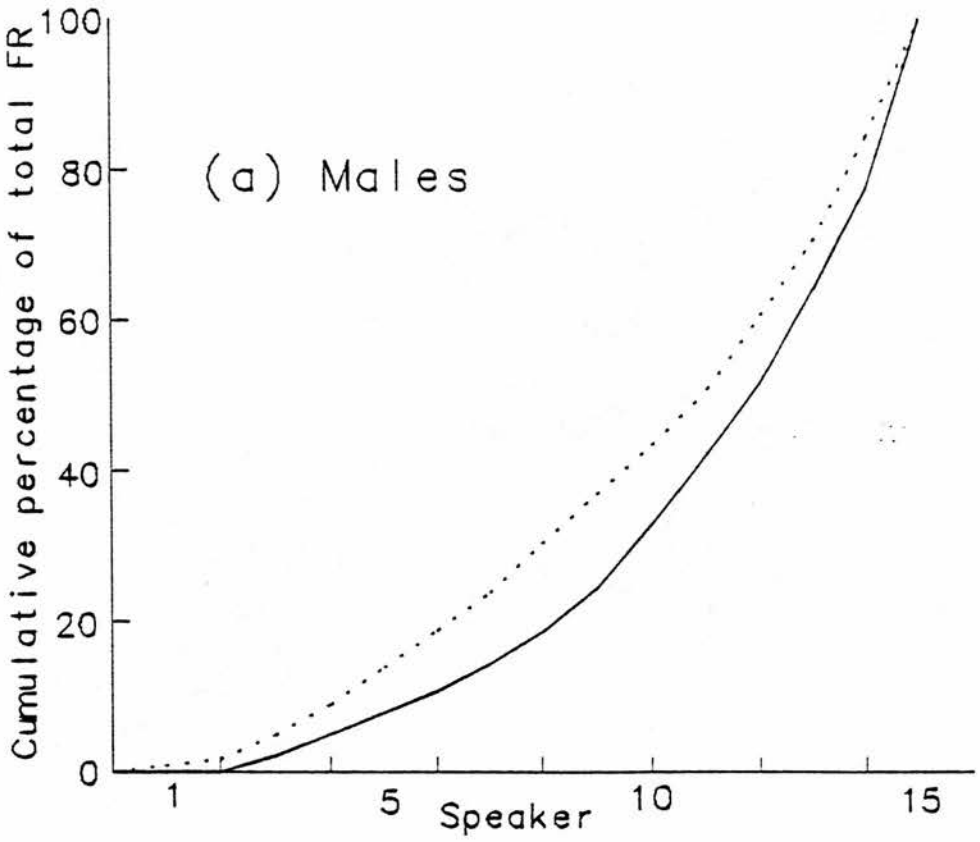

Figure 7.9 Cumulative percentage of total False Rejection errors by speaker, using global (solid line) and speaker-specific (dotted line) thresholds. Orthogonalized spectral features, correlation classifier.

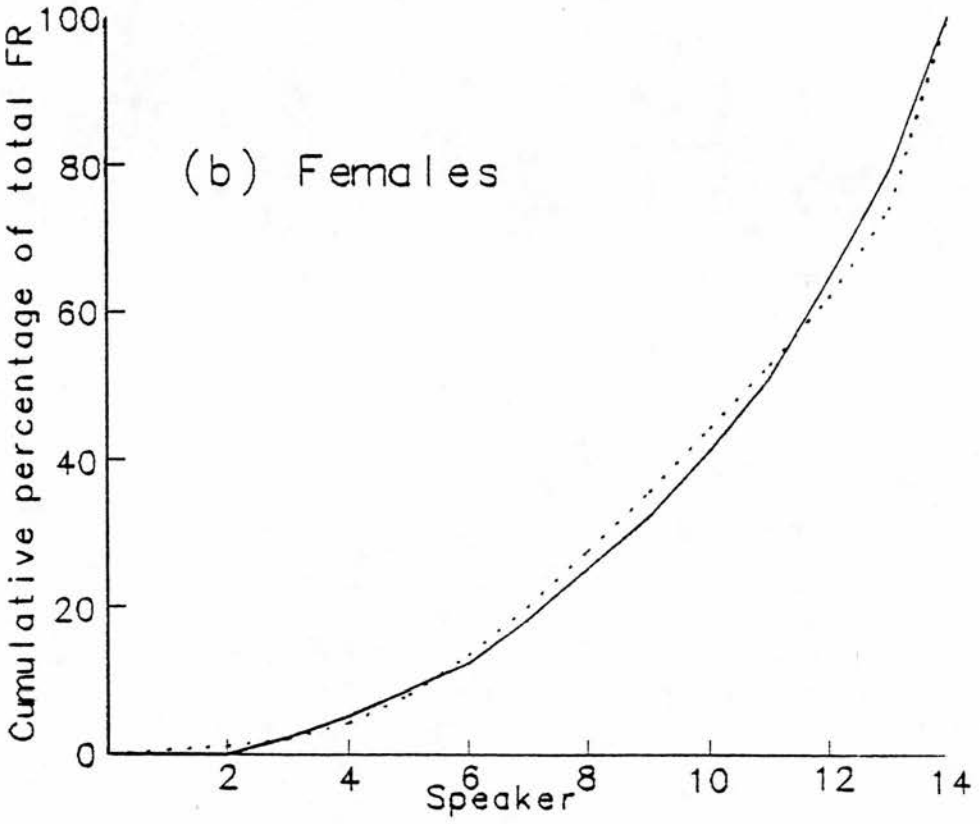

Figure 7.9 Cumulative percentage of total False Rejection errors by speaker, using global (solid line) and speaker-specific (dotted line) thresholds. Orthogonalized spectral features, correlation classifier.

It can be seen that with the global threshold (solid line), over half of the FR errors are accounted for by only the top three speakers – less than 25 % of the population – for both male and female groups. The cumulative density function for the speaker-specific thresholds (dotted line) shows that the use of separate thresholds has spread the number of errors a little more widely, but the essential shape of the distribution remains the same, with only a small number of speakers accounting for over half the total number of errors.

7.6.4. Discussion

These experiments confirm that the inhomogeneity of speaker groups is a major problem for speaker verification systems (Doddington 1985). The use of speaker-dependent thresholds is only a partial solution, however, and may actually make the system more vulnerable to impostor attempts, since certain speakers may have such wide thresholds that they pose a security risk.

These experiments also illustrate the problems that would be faced if one were trying to optimise a system using a relatively small database. It is quite possible that a smaller database than the one used here would have excluded at least some of the worst-performing speakers, giving a much better (but false) picture of the system's future performance. Equally, a larger database might reduce the significance of these speakers' contribution to the overall error rates, without including more unstable speakers.

7.7. Distribution of errors over time: the need for adaptation

7.7.1. Introduction

The performance of a speaker verification system over time is likely to be affected by several factors. One is that speakers' voice characteristics vary from day to day, and even at various times during the same day. In addition, there are gradual changes over time which may have a physiological basis (ageing or disease, for example) or a sociolinguistic one (accommodation to a new accent group). A third factor influencing performance is that speakers' attitudes to the use of an automatic system tend to change over time. It has been observed in working systems that the rate of False Rejections is usually at its highest in the first few sessions after enrolment, when speakers feel intimidated by the system (Doddington 1985), and perhaps too self-conscious. As their experience of the system grows, however, their performance stabilises, and FR rates fall to a lower level.

These factors mean that a reference profile derived on one day may perform badly on speech material gathered some time later. One solution to this problem – the solution adopted so far in this thesis – is to derive the reference from speech material gathered over several days, in the hope of including more of the speakers' long-term distributions within the reference. Such a strategy can only hope to cope with some of the shorter-term variation, however, since the longer-term physiological or sociolinguistic changes are unlikely to be sampled over periods of just a few weeks. In addition, an extended enrolment of this sort is not always possible (e.g. in telephone banking applications), and

references must be derived from only one or two sessions of data. In these circumstances, some form of reference *adaptation* may be desirable.

Adaptation of reference information may take several forms. At its simplest, information from a successful bid is included in the reference itself (by averaging the two vectors, for example), so that the reference gradually accommodates to any changes in the speaker's performance. At the same time, the thresholds used to judge the bids may be revised, as the system's knowledge of intra-speaker and inter-speaker distance distributions increases. At its most sophisticated, adaptation may also include a re-appraisal of the features chosen for verification, to maintain optimal performance.

Adaptation may therefore be an answer both to the problems of (permanent) changes in speakers' performance over the longer term, and to the need to obtain a large amount of training data to give stable references in the short term, since the inclusion of extra bids effectively prolongs the enrolment phase indefinitely.

7.7.2. Intra-speaker distances over time

To gain some idea of whether adaptation might improve the performance of the nasal spectral vectors, an experiment was carried out to measure the average intra-speaker distance as a function of time after enrolment, as suggested by Rosenberg and Shipley (1983). In the absence of profile adaptation, it is the changes in intra-speaker distance and therefore the rate of False Rejection which will determine the changes in error rate, since it is unlikely that impostor bids will show any similar sort of trend. Rosenberg and Shipley found

that intra-speaker distances (as indicated by the mean) rose quite sharply over the three sessions following enrolment, requiring an increase in acceptance thresholds to maintain a constant rate of False Rejection, but thereby allowing a greater degree of False Acceptance too.

Method

In this experiment, the number of sessions over which distances were measured was increased from four to six, by deriving a new set of references for the 15 male and 14 female speakers using only their first *two* sessions; tokens from the remaining six sessions were therefore available for use as bids. The references were formed by averaging the tokens from the first two sessions for each speaker, and projecting the mean vector for that speaker on to the orthogonalised feature space using the canonical discriminant functions derived in 7.5. Test bids were also projected in the same way, and bids were made against the references using both the Euclidean distance classifier and the correlation classifier. The changes over time were tracked by pooling the distributions of the resulting intra-speaker distances over all speakers within each sex group for each session. The median was chosen as a suitable summariser, rather than the mean (as in Rosenberg and Shipley 1983), because the choice of a threshold for speaker verification is made on the basis of the cumulative frequency distributions.

Results

Figures 7.10 (a) and 7.10 (b) show the variation in the median intra-speaker distances for the male and female speakers respectively, for both classifiers. To allow the distances from each classifier to be compared directly, the median values were standardized by the mean median value in each case.

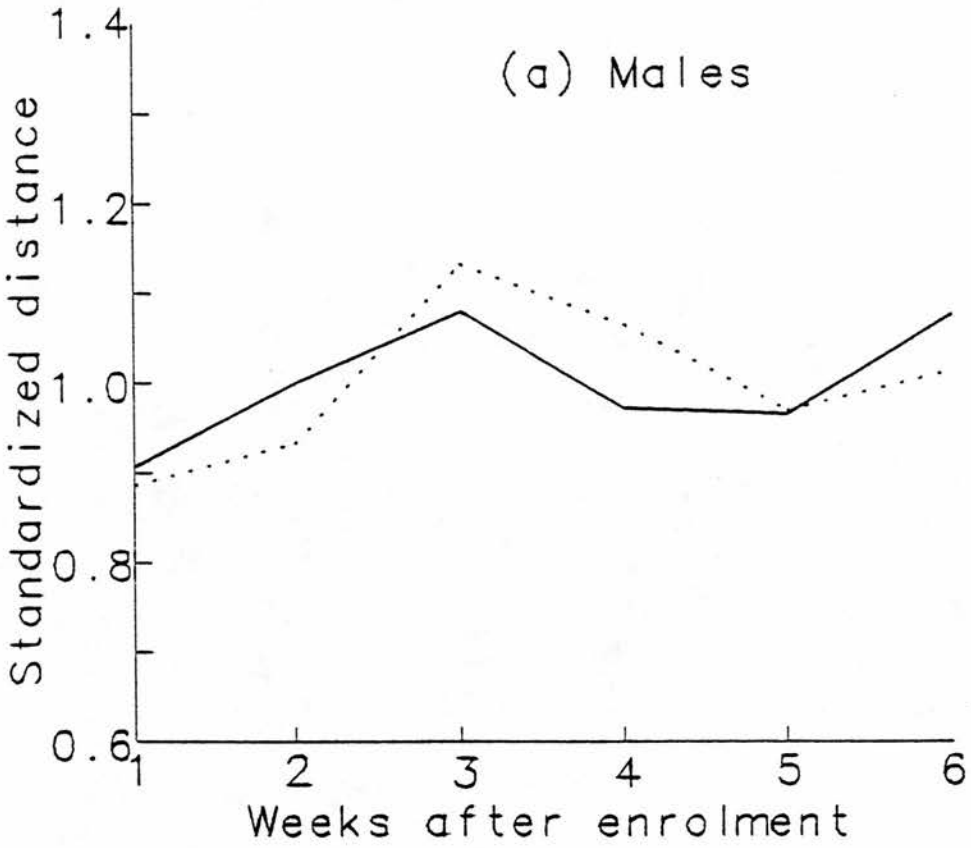

Figure 7.10 Standardized median intra-speaker distances by week for (a) 15 male and (b) 14 female speakers; orthogonalized spectral vectors, with Euclidean (solid line) and correlation (dotted line) classifiers

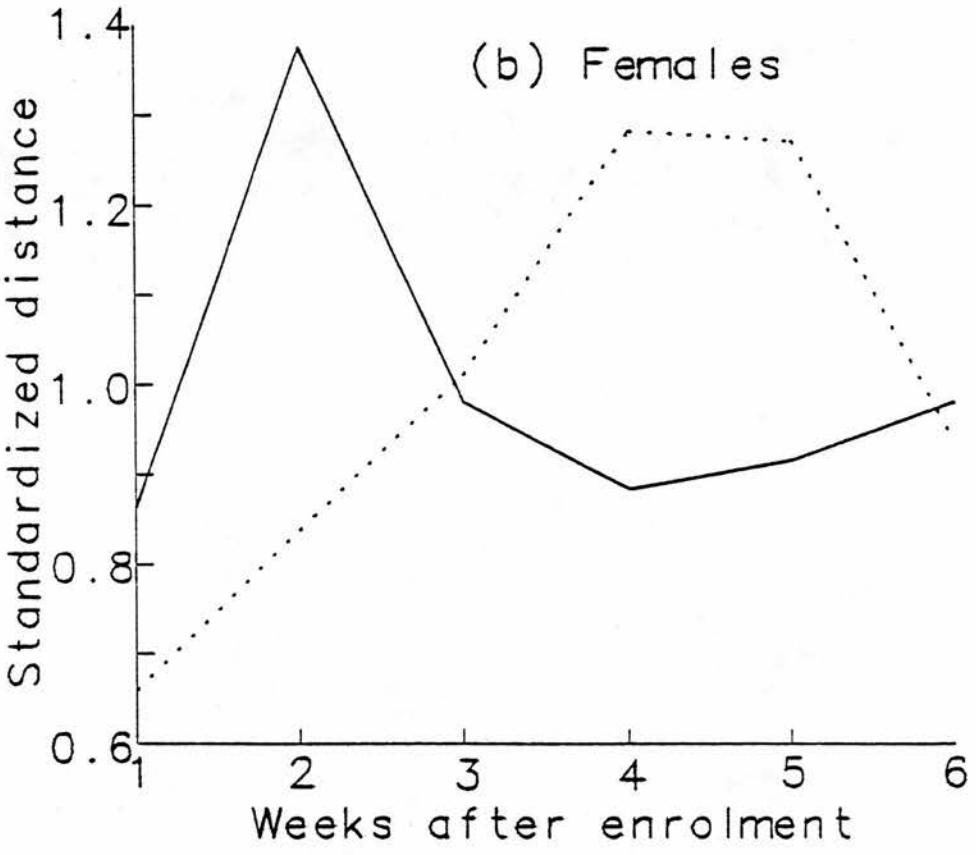

Figure 7.10 Standardized median intra-speaker distances by week for (a) 15 male and (b) 14 female speakers; orthogonalized spectral vectors, with Euclidean (solid line) and correlation (dotted line) classifiers

Discussion

There is clearly considerable variation from week to week in these distances, and some difference between the performance of the two classifiers. The male speakers seem to show a small increase with time, but there is not the pronounced upward trend observed by Rosenberg and Shipley (1983). There are two possible reasons for this. The first is that since the references were derived from data spread over two weeks, a considerable amount of the variability of each speaker may already have been sampled. The second explanation is that there is no reason to expect speakers' distances to go on rising with increased separation from their enrolment: the differences in speakers' performance over periods of a few weeks (or perhaps months) are random in nature, since it is unlikely that any longer-term changes take place in this time. Unless the choice of reference utterances was particularly ill-timed, therefore, their utterances are just as likely to move closer to their reference profiles as further away.

7.7.3. Adaptation strategies

While adaptation may not strictly be necessary in this case, it may nevertheless be beneficial to performance, by making the references even more robust, and also by coping with non-random changes over the short term, such as those caused by health problems.

The simplest method of adaptation consists in averaging each successful bid made against speaker's reference with the reference itself. Thus the reference can shift with changes in the speaker's distribution. A straightforward

averaging of the two vectors gives a disproportionate weight to the bid vector, however, a weight which it would not have if it had simply been included as one of the reference vectors used to derive the mean. The bid reference vector must therefore be given a reduced weighting in the calculation. The default weighting, which gives equal weight to all vectors included in the reference, is:

$$\frac{n}{(n+1)} * r \; + \; \frac{1}{(n+1)} * b \tag{7.5}$$

where r is the reference vector, b is the bid vector and n is the number of vectors included so far in the formation of the reference. The relationship between the two weighting coefficients is thus defined as

$$w_r = 1 - w_b \tag{7.6}$$

where $w_b = 1/(n+1)$. As the number of vectors included in the reference increases, therefore, the weight on the bid is reduced.

This strategy carries with it the risk of contamination of the reference by impostor bids, however, since the only check on the identity of the speaker is that provided by the system itself. Impostor bids which are accepted will be incorporated into the reference, making subsequent impostor bids more likely to succeed (increasing the False Acceptance rate) and possibly reducing the benefits of adaptation for the False Rejection rate too by moving the reference away from the true centroid for that speaker.

One way of reducing this risk is to set an additional threshold, lower than that for simple acceptance of the bid, which bids must not exceed if they are to be accepted for adaptation. This could be set to exclude the lowest inter-speaker distance, for example. A more sensitive alternative is to use an

additional weighting factor on the bid, inversely proportional to the distance achieved. This does away with the need to estimate an explicit threshold. The *certainty score* associated with each bid (section 7.3) is one suitable weighting factor: it expresses the probability that the bid distance obtained comes from the intra-speaker distribution, on a scale between 0 % (definitely not genuine) and 100 % (certainly genuine). The weighting on the bid would therefore be

$$w_b = \frac{c}{(n+1)} \tag{7.7}$$

where c is the certainty score (Equation 7.3, expressed here as a fraction between 0 and 1 rather than as a percentage) for bid b, with the reference weight w_r equal to $1-w_b$ as before. Thus the less certain it is that the bid, though accepted,comes from the intra-speaker distribution, the less weight that bid receives when included in the reference. Only when the system was absolutely certain that the bid came from a genuine speaker would the bid receive equal weight with the training vectors, as in Equation 7.4.

7.7.4. An experiment in reference profile adaptation

The strategies outlined above were implemented in an experiment using the 15 male and 14 female speakers. In this experiment, the empirical False Rejection and False Acceptance rates were calculated in trials a) without adaptation of reference profiles b) with adaptation taking place on successful verifications and c) with adaptation weighted by the certainty score for successful bids. Vowel-specific references were used, and the classifier used the corre-

lation measure.

Method

The acceptance threshold for verification was set at the Equal Error Rate threshold estimated during a modified training phase. The first two sessions for each speaker were used to form references, with the canonical discriminant functions based on the same two sessions being used to project them on to the new feature space. The third and fourth sessions for each speaker were then used to calculate the *a-posteriori* Equal Error Rate and the certainty function for each sex, for each vowel context (based on the intra-speaker and inter-speaker distance distributions). The distance thresholds for each vowel context (global thresholds for each sex) were set at the Equal Error Rate threshold for that context.

The bid data came from the last four sessions of each speaker as in section 7.5, but were reduced in number. In the experiments reported there, the "impostor" bids greatly outnumbered the genuine bids (by as much as 14 to 1). The use of the Equal Error Rate Threshold meant that, while the probabilities of False Rejection and False Acceptance would be roughly equal, the *number* of impostor bids accepted would be likely to exceed the number of genuine bids for any given speaker, putting the system at an unrealistic disadvantage, since in real life it is unlikely that impostor bids would outnumber genuine ones. In this experiment, therefore, the number of "impostor" bids against each speaker was limited to 36, half the number of available genuine bids, giving a total of 108 bids per speaker. These impostor bids were randomly selected from among

the remaining speakers. Impostor and genuine bids were then interleaved in a random order (the same for each speaker) so that contamination of references by impostors was allowed in as realistic a way as possible.

Results

The resulting False Rejection and False Acceptance error rates are shown in Table 7.9, for the two adaptation strategies and the absence of adaptation. The rates without adaptation are higher than those achieved in section 7.5, but the two experiments are not really comparable since the numbers of bids differ greatly.

Discussion

Speaker group	Error rate	Adaptation strategy		
		a	b	c
Males (15)	FR	15.31	7.98	10.11
	FA	15.74	21.30	19.07
	AVE	15.53	14.64	14.59
Females (14)	FR	26.09	12.60	16.67
	FA	12.10	14.09	13.69
	AVE	19.10	13.35	15.18

Table 7.9 Error rates (%) for 14-dimensional (male) and 13-dimensional (female) orthogonalized spectral vectors using EER threshold, for three adaptation strategies: a) no adaptation b) unweighted adaptation c) weighted adaptation (certainty score)

For both sexes, the use of an adaptation strategy has significantly reduced the False Rejection rate, as might be expected. The rates of False Acceptance have risen, indicating that speakers' references are indeed becoming contaminated by impostor bids. The overall error rate, given here by the *average* of the FR and FA rates, has fallen, however, suggesting that adaptation has indeed been worthwhile.

The two adaptation strategies do not differ markedly in their effect on the average error rate, but their effects on the separate FA and FR rates are different. The use of equal weighting for the bid and reference vectors (strategy (b), Equation 7.1) produces the lower FR rate, with a correspondingly higher FA rate. The use of the certainty score as an additional weighting coefficient on the bid (strategy (c), Equation 7.7) gives a higher FR rate (though still much lower than that obtained without adaptation), while lowering the FA rate somewhat.

Neither of the strategies used here adapts the distance *thresholds*, though this has been found useful in some studies (e.g. Rosenberg and Sambur 1983, Fakotakis et al. 1987). Fakotakis and co-workers suggest that it is the distance threshold adaptation which actually gives the greatest benefit. It is possible, therefore, that an even greater improvement in performance could be achieved if this were included. Possible strategies, other than a complete re-appraisal of the intra- and inter-speaker distance distributions at each adaptation (which would be extremely costly) include estimating a constant compensation factor during training, to be applied to the threshold each time adaptation takes

place; and a system whereby the change in the most recent bid distances as a result of reference adaptation is estimated, and used to modify the thresholds proportionately.

7.8. Summary and discussion

This chapter has examined the performance of a set of features extracted from tokens of the velar nasal stop. In studies of the choice of feature set, it was found that the raw pole-zero spectrum performed better than peak and dip features extracted from the separate all-pole and all-zero spectrum, despite the promise shown by these peak-dip features in the F-ratio analyses of Chapter Six and the evidence in this chapter (section 7.5) that the discrimination information in the spectrum is concentrated in particular frequency locations. The failure of the peak-dip features to perform as well as expected is thought to be due partly to the difficulties experienced in obtaining a uniform number of peaks and dips from tokens by the same speaker. The distance measure adopted – the warped Euclidean distance presented in Chapter Six – appears to be unable to overcome this problem, and may in fact have exacerbated it by giving equal or greater weight to such differences between profiles as to the differences in the peak-dip frequencies themselves.

The pole-zero spectral vectors were found to give a reasonably good performance, though this is not as good as that obtained using vowel-based parameters elsewhere in the literature (e.g. Feix and De George 1985, Doddington 1985). It was also found that they responded well to various optimisation procedures, such as the use of canonical analysis for data reduction and

elimination of redundancy (7.5), and the introduction of an adaptation strategy (7.7).

Vowel context was found to have a significant effect on performance, despite the reported resistance of the velar nasal's characteristics to coarticulation (Chapter Three). This appears to vindicate the view expressed in Chapter Three that the stability of nasal stops in general tends to be overrated. Large differences found between speakers in error rates could perhaps also be attributed to such short-term instability: speakers are known to vary considerably in the extent to which coarticulation with vowels takes place in nasal stops (Su et al. 1974).

Direct comparison with existing studies using nasality has been avoided in this chapter, because it is extremely difficult: the results achieved depend not only on the feature set employed, but on the nature of the classifier, the application of statistical optimisation and the speaker database itself, as was made clear by the later sections of this chapter. In addition, much of the work on nasality has used the *identification* task (e.g. Glenn and Kleiner 1968), which increases the difficulty of comparing results, because of the different way in which performance is measured.

Factors which were not explored in this chapter include the effects of health changes during the months in which the database was recorded, and the performance of the nasal spectra on proper impostor bids. A study of the effects of speakers' health would require detailed knowledge of their physiological state at the time of each recording, and this knowledge was not available.

Such a study might also be difficult to interpret, given the variability observed in the current database of apparently healthy speakers. It would, however, prove useful in designing ways of compensating for the changes known to take place. The use of "casual" impostor bids would also have been desirable, given the twelve-fold increase in errors reported by Lummis and Rosenberg (1972), using Doddington's (1971) feature set. However, it would be difficult to achieve a fair test of speakers' abilities at imitating nasal spectra. It is likely that their attempts at imitation would have focussed on phonatory parameters, durations and vowel qualities – features which are at the heart of systems such as Doddington's, and those which use Dynamic Time Warping of spectral or cepstral information. Thus we would not actually be testing the resistance of the nasal spectrum itself to impersonation (except as far as coarticulation effects from neighbouring vowels were concerned) – only showing that its subtlety makes effective impersonation less likely.

CHAPTER EIGHT

SUMMARY AND CONCLUSIONS

CHAPTER EIGHT

SUMMARY AND CONCLUSIONS

8.1. Introduction

This chapter provides a brief summary of the main findings of the work reported in earlier chapters, makes some suggestions for further work in this area and considers some of the issues not yet addressed.

8.2. Summary of the preceding chapters

This thesis has considered the use of nasality in Automatic Speaker Verification by focussing on the potential of a single segment type, the velar nasal stop [ng]. The main themes of this thesis have been the choice of a suitable manifestation of nasality, the choice of an appropriate method of characterising it acoustically, and an exploration of the nature and extent of the variability in the nasal spectrum, through descriptive statistical analysis and simulated speaker verification trials.

The choice of velar nasal stops was made after extensive reviews of both the speaker verification field (Chapter Two) and the nature of nasality (Chapter Three). In Chapter Two it was argued that a segmental approach to verification, despite the need for segment location, could be beneficial because of its reduced dependence on text and the small amount of speech needed. The

desirability of using speech features with maximal dependence on the physical basis of speech was highlighted: such features are likely to show fairly large differences between speakers, and are generally difficult to modify voluntarily. In Chapter Three, the physical, phonetic and acoustic characteristics of nasality in all its forms were reviewed, and the dependence of the nasal spectrum on relatively fixed anatomical features was confirmed. It was also noted, however, that nasality does not offer an invariant acoustic marker of identity: the nasal cavities themselves are still subject to physiological change (often fairly rapid), while both passive and voluntary movements elsewhere in the vocal tract have a significant effect on the nasal spectrum. Velar nasal stops were found to offer the best phonetic environment for the use of nasality: they show the greatest dependence on the nasal tract for their acoustic characteristics, and the greatest resistance to external effects such as lingual coarticulation.

The *measurement* problems posed by the use of nasality are indeed considerable. In this thesis, they were addressed by the choice of a method capable of accurate location of both spectral poles and zeros, without the distortion introduced by methods such as Linear Prediction. The pole-zero decomposition technique (Yegnanarayana 1981) introduced in Chapter Four proved quite suitable for this, giving a representation which allowed the separate measurement of pole (peak) and zero (dip) frequencies using a simple peak-picker. Application of the method to tokens of all three British English nasal stops gave results comparable with the limited data published in the literature (Chapter Five). The use of a higher model order for the all-pole model than for the all-zero

model was found to be necessary, and follows from the nature of the speech production mechanism.

The nature of the variability of the peak and dip features of the velar nasal was explored in Chapter Six. The extent of this variability in the velar nasal had not previously been studied in a large group of speakers. The variability in the *numbers* of peaks and dips located in the spectrum necessitated the use of a technique for restoring the proper alignment of peak and dip features in speakers' profiles before a statistical analysis was undertaken of the main trends (sex, vowel context, speaker differences and intra-speaker consistency over time). This analysis confirmed that speakers' identity was the dominant factor in accounting for the variability in peak and dip frequencies, that vowel context effects were minimal but still present, and that intra-speaker variation was considerable. The spectral dips proved to be less reliable in their occurrence (many tokens had none) and also to have lower F-ratios than the peak features, suggesting that they would be of limited use for speaker verification on their own. The use of the pole-zero decomposition technique is still recommended, however, because it gives an inherently more accurate analysis of the *peak* frequencies, and of the combined pole-zero spectrum.

The use of the spectral features of the velar nasal for speaker verification was explored in Chapter Seven. The poor performance of the peak and dip features, partly attributable to the problems of obtaining a consistent number of elements from speakers' tokens, led to the choice of the combined pole-zero spectrum as the more suitable representation. The use of context-dependent

reference information, the introduction of variance-based feature weighting in the calculation of distances, and the application of canonical analysis for the purposes of orthogonalization and dimension reduction, all proved beneficial to varying degrees. Equal Error Rates of 12.8% for 15 male speakers and 13.8% for 14 female speakers were achieved, demonstrating that a reasonable level of performance is possible using velar nasal stops. The use of a simple reference profile adaptation strategy also helped to reduce the error rates observed.

8.3. Areas for further work

Some possible areas for improvement and investigation have already been indicated in the preceding chapters, and are summarised here.

The need for reliable peak-dip detection causes considerable difficulties in many phonetic studies. The problem of identifying missing and spurious peaks was approached here using the technique of *peak profile warping*. This technique gave a reasonably good alignment, but errors still occurred. Further development might prove useful in other studies of this type, where there is a need to identify formant peaks automatically. The choice of the prototype is one area which could be made more robust: various other methods could be investigated, such as the "minimax" technique (choose the vector with the smallest maximum distance from all other vectors), and many forms of cluster analysis, which might lead to the choice of more than one prototype.

A revised method of calculating the overall distance may also be required if the technique is to be used for cross-speaker comparisons in automatic speaker verification. For example, it may be desirable to give greater weight to

differences in the *frequency* of corresponding peaks than to differences in the *number* of peaks, since speakers show so much internal variability. Alternatively, a method of peak detection which does away with the need for peak warping by guaranteeing a fixed number of peaks might prove useful. However, the use of such a method will almost certainly lead to a loss of sensitivity to genuine differences.

It was suggested in Chapter Six that a more controlled study of vowel context effects, using perhaps a smaller number of speakers and a greater number of tokens recorded in a single session, would shed more light on the extent of coarticulation. The requirements of such a study are in some ways incompatible with those of speaker verification, in which a sample of the full range of variation in voice characteristics is desirable.

Finally, the experiments reported in Chapter Seven show how the performance of a set of features depends heavily on the classifier used and statistical optimisation employed. Further improvements to the error rates obtained would be possible using additional techniques such as adaptation of distance thresholds, taking advantage of the reduction in intra-speaker distance resulting from the reference adaptation to reduce the False Acceptance rate at the same time.

8.4. Some outstanding issues

Three major areas have not been addressed so far in this thesis. These are the need for segmentation, the effects of physiological change on the nasal spectrum, and the possible effects of impersonation.

8.4.1. The need for automatic segmentation

In line with most studies on the use of nasal segments for speaker verification or identification, this study has used nasal segments isolated by hand. To be truly automatic, however, automatic location of the segments of interest is required. This has not been attempted in this thesis, since it introduces another possible source of error which can obscure the potential of the nasal segments. This is not a trivial problem, however, since segmentation errors can contribute significantly to the rejection rate of a system (Das and Mohn 1971).

There is not yet a fully reliable automatic nasal detection algorithm. The best performance on continuous speech has been achieved using a two-stage decision process, in which possible nasal stops are located using gross spectral energy parameters such as the first spectral moment (the energy centroid of the spectrum: Williams et al. 1989) or the presence of energy dips in particular spectral bands (Weinstein et al. 1975), and then classified as nasal/non-nasal (often with place of articulation classification too) by means of measures such as segment durations, formant frequencies, bandwidths and relative amplitudes, or the presence of a low frequency resonance. The use of spectral energy tends to be fairly good at locating nasal stops (approximately 95% of nasal stops were found by Glass and Zue (1986), and by Williams et al. (1989)), but it also gives a very high proportion of "false alarms" – that is, non-nasal segments incorrectly labelled as nasal: 67% of the total number of segments found by Glass and Zue (1986) were non-nasal, while for Williams et al. the figure was

43%. The use of the second stage helps to eliminate many of these "impostor" nasals (an average of 3.5% remained in Williams et al. (1989), and 15% in Glass and Zue (1986)). However, correct recognition of true nasals is poorer at this stage, with false rejection rates of between 21% (Glass and Zue 1986) and 36% (Williams et al. 1989).

While these results seem reasonably good, it appears that many of the errors occur either with the *velar* nasal (Williams et al. 1989) or with postvocalic nasals in general, because they tend to be shortened or elided altogether (Weinstein et al. 1975). Another problem with several of these algorithms is that, being designed for speech recognition applications, they are biassed towards the recognition of nasal stop *phonemes*, even when their phonetic exponents – nasal stop phones – are missing (e.g. Glass and Zue 1985, 1986). This is not satisfactory for speaker verification, since actual tokens of the nasal stop are required.

The problems of nasal detection are lessened slightly in speaker verification where a fixed text can be used, since the knowledge of the segmental content of the utterances to be spoken can be used to guide boundary location and segment identification, assuming that speakers are cooperative.

8.4.2. Physiological change

There is very little information on the exact nature of the acoustic variability resulting from physiological changes such as those accompanying head-colds and the like. This thesis has not addressed the problem directly, since the factors underlying the variability from week to week in the nasal resonance

patterns were not explored. This is because such an exploration, to be of any use, would have to have some way of calibrating the *physiological* change as well as the *acoustic* change. Even with detailed knowledge of the physiological changes, however, the extent of the variability observed in Chapter Six would make interpretation of any acoustic changes rather difficult. If these problems could be overcome, however, more detailed knowledge of how health problems affected the nasal spectrum might allow some form of compensation, or rapid short-term adaptation, to be introduced, making the use of nasal spectral features more robust.

8.4.3. Impersonation: the use of trained impostors

The experiments reported in Chapter Seven used "casual" impostors throughout. It is essential, however, to know how far determined mimics can imitate the nasal resonance patterns of other speakers. Presumably the only way in which they could do this is by compensatory articulations elsewhere in the vocal tract, such as by the use of larynx lowering to lower all resonance frequencies and decrease the spacing between resonance peaks (though the effects of this on the nasal spectrum have not been studied: see Chapter Three, 3.6.2). Alternatively, imitations of speakers' *vowel* quality might help to bring their nasal spectrum closer to that of the target speaker, though the scope for this is limited in the case of the velar nasal, as shown in Chapter Six. A certain amount of protection is given by the lack of salience of the nasal spectrum, but it might still be possible for voluntary alterations to a speaker's vocal tract – even if not specifically aimed at recreating a target speaker's nasal spectrum –

to succeed in getting close enough to be accepted, given the amount of intra-speaker variability observed in this thesis. However, if nasality were one feature in a multi-processor system, it is likely that even in this eventuality their attempt would fail because the effects of compensatory gestures would be seen in other aspects of speech production (a lowered fundamental frequency and increased breathiness, for example, in the case of larynx lowering).

APPENDICES

APPENDIX A

MRPA – MACHINE-READABLE PHONEMIC ALPHABET

This thesis uses the Machine Readable Phonemic Alphabet in place of the usual IPA symbols, to represent both phonemic and phonetic categories. The principal correspondences between MRPA and the IPA alphabet are given below.

MRPA	IPA	MRPA	IPA
i	ɪ	p	p
ii	i	t	t
e	e	k	k
a	æ	b	b
aa	ɑ	d	d
uh	ʌ	g	g
@@	ɜ	m	m
@	ə	n	n
o	ɒ	ng	ŋ
oo	ɔ	h	h
u	ʊ	f	f
uu	u	th	θ
ei	eɪ	s	s
ou	əʊ	sh	ʃ
au	aʊ	v	v
ai	aɪ	dh	ð
oi	ɔɪ	z	z
i@	ɪə	zh	ʒ
u@	ʊə	ch	tʃ
e@	ɛə	jh	dʒ
w	w	l	l
y	j	?	ʔ
r	ɹ		

APPENDIX B

WORDS USED IN DATABASE RECORDINGS

This is a list of the words used in experiments in Chapters Six and Seven.

For sessions Two to Eight, only those words shown in bold were used.

1. Session One

sung	**hung**	**rung**	**bung**	**tongue**
fang	**bang**	**rang**	**hang**	**sang**
ring	**king**	**sting**	**sing**	**wing**

2. Sessions Two to Eight

Tam	can	Pam	can
Pam	kin	**ting**	twin
ping	queue	Kim	Pam
ting	twin	**pang**	pin
pin	pure	**tongue**	**pang**
Kim	Tam	**king**	pun
tang	**ping**	spine	spine
pang	pin	can	kin
tin	**tang**	queue	pure
tongue	tin	pure	ping
pun	pun	Tam	Kim
king	**stung**	**ting**	tin
stung	scan	**tang**	king
spine	kin	**tongue**	scan
scan	twin	**stung**	queue

BIBLIOGRAPHY

BIBLIOGRAPHY

Abercrombie, D. (1967), *Elements of General Phonetics,* Edinburgh University Press, Edinburgh.

Abramson, A.S., Nye, P.W., Henderson, J.B., and Marshall, C.W. (1981), "Vowel height and the perception of consonantal nasality", *J. Acoust. Soc. Amer.*, vol. 70, pp. 329-338.

Ashby, M.G. (1983), "Effects of variation in larynx height", *Speech, Hearing and Language: work in progress, University College London*, vol. 1, pp. 29-39.

Atal, B.S. and Hanauer, S.L. (1971), "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Amer.*, vol. 50, pp. 637-655.

Atal, B.S. (1972), "Automatic Speaker Recognition Based on Pitch Contours", *J. Acoust. Soc. Amer.*, vol. 52, no. 6 (part 2), pp. 1687-1697.

Atal, B.S. (1974), "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304-1312.

Atal, B.S. (1976), "Automatic Recognition of Speakers from Their Voices", *Proc. IEEE*, vol. 64, no. 4, pp. 460-474.

Atal, B.S. and Schroeder, M.R. (1978), "Linear prediction analysis of speech based on a pole-zero representation", *J. Acoust. Soc. Amer.*, vol. 64, pp. 1310-1318.

Atal, B.S. (1985), "Linear predictive coding of speech", in *Computer Speech Processing*, ed. F. Fallside and W.A. Woods, pp. 81-124.

Beddor, P.S. (1983), *Phonological and phonetic effects of nasalization on vowel height,* Indiana University Linguistics Club, Bloomington, Indiana.

Bell-Berti, F., Baer, T., Harris, K.S., and Niimi, S. (1979), "Coarticulatory effects of vowel quality on velar function", *Phonetica*, vol. 36, pp. 187-193.

Bell-Berti, F. (1980), "Velopharyngeal function: a spatial-temporal model", in *Speech and Language: Advances in Basic Research and Practice*, ed. N.J. Lass, vol. 4, pp. 291-316.

Berg, J. van den (1962), "Modern research in experimental phoniatrics", *Folia Phoniatrica*, vol. 14, pp. 81-149.

Bielby, G., Lennig, M., and Mermelstein, P. (1987), "Speaker verification with sequential decision on a speaker specific vocabulary", Presented at European Conference on Speech Technology, Edinburgh 1987.

Bjuggren, G. and Fant, G. (1964), "The nasal cavity structures", *STL-QPSR-4/1964*, pp. 5-7.

Bogner, R.E. (1981), "On Talker Verification Via Orthogonal Parameters ", *IEEE Trans. Acoust. Speech + Sig. Proc.* , vol. 29, no. 1, pp. 1-12.

Bozic, S.M., *Digital and Kalman Filtering*, 1979.

Brown, R. (1980), "Auditory Speaker Recognition: a theoretical and experimental study", PhD thesis, University of Edinburgh.

Buck, J.T., Burton, D.K., and Shore, J.E. (1985), "Text-Dependent Speaker Recognition Using Vector Quantization", *Proc. IEEE ICASSP-85*, vol. 1, pp. 391-394.

Bunge, E. (1977), "Automatic Speaker Recognition System AUROS for Security Systems and Forensic Voice Identification", *Proc. 1977 Internat. Conf. on Crime Countermeasures - Sci. + Eng.*, pp. 1-7.

Cagliari, L.C. (1978), "An experimental study of nasality with particular reference to Brazilian Portuguese", Ph.D. thesis, University of Edinburgh.

Calnan, J. (1955), "Movements of the soft palate", *Speech*, vol. 154, pp. 14-20.

Carr, P.B. and Trill, D. (1964), "Long-term larynx-excitation spectra", *J. Acoust. Soc. Amer.*, vol. 36, pp. 2033-.

Castelli, E. and Badin, P. (1988), "Vocal tract transfer function measurements with white noise excitation application to the naso-pharyngeal tract", *Proc. 7th FASE Symposium, Edinburgh 1988*, pp. 415-422.

Catford, J.C. (1977), *Fundamental Problems in Phonetics*, Edinburgh University Press, Edinburgh.

Cheung, R.S. (1978), "Feature selection using adaptive learning networks for text-independent speaker verification", *J. Acoust. Soc. Amer.*, vol. 64 Suppl.1, p. S183 (A).

Cheung, R.S. and Eisenstein, B.A. (1978), "Feature Selection via Dynamic Programming for Text-Independent Speaker Identification", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-26, no. 5, pp. 397-403.

Chen, S. and Lin, M. (1987), "On the use of pitch contour of Mandarin speech in Text-Independent Speaker Identification", *Proc. IEEE ICASSP-87*, vol. 3, pp. 1418-1421.

Childers, D.G., Skinner, D.P., and Kemerait, R.C. (1977), "The cepstrum: a guide to processing", *Proc. IEEE*, vol. 65, pp. 1428-1443.

Clarke, F.R. and Becker, R.W. (1969), "Comparison of techniques for discriminating among talkers", *J. Speech and Hearing Research*, vol. 12, pp. 747-761.

Corsi, P. (1981), "Speaker recognition: a survey", in *Automatic Speech Analysis and Recognition*, ed. Jean-Paul Haton, D. Reidel Publishing Company, Dordrecht, Boston, London.

Cox, R.C. and Robinson, D.M. (1980), "Some notes on phase in speech signals", *Proc. IEEE ICASSP-80*.

Crothers, J. (1975), "Nasal consonant systems", in *Nasalfest: Papers from a Symposium on Nasals and Nasalization*, ed. C.A. Ferguson, L.M. Hyman and J.J. Ohala, pp. 153-166.

Crowe, A. and Jack, M.A. (1987), "A globally optimising formant tracker using generalised centroids", *Electronics Letters*, vol. 23, pp. 1019-1020.

Crystal, D. (1969), *Prosodic Systems and Intonation in English,* Cambridge University Press, London.

Czermak, J.N. (1857), "Uber das Verhalten des Weichen Gaumens beim Hervorbringen der reinen Vocale", *Wiener Akademie Sitzungsberichte*, vol. 27.

Czermak, J.N. (1858), "Uber reine und nasalierte Vocale", *Wiener Akademie Sitzungsberichte*, vol. 28.

Czermak, J.N. (1869), "Wesen und Bildung der Stimmund Sprachlaute", in *Czermak's gesammelte Schriften (Vol.2)*, Engelmann, Leipzig.

Das, S.K. and Mohn, W.S (1971), "A Scheme for Speech Processing in Automatic Speaker Verification", *IEEE Trans. on Audio and Electroacoustics*, vol. AU-19, no. 1, pp. 32-43.

Das, S.K., Mohn, W.S, and Saleeby, S.L. (1971), "Speaker verification experiments", *J. Acoust. Soc. Amer.*, vol. 49, p. 138(A).

Delattre, P. (1954), "Les attributs acoustiques de la nasalite vocalique et consonantique", *Studia Linguistica*, vol. 8, pp. 103-109.

Delattre, P. (1969), "Two types of nasality: vocalic and consonantal", in *The General Phonetic Characteristics of Languages*, pp. 81-100, U.S. Dept. of Health, Education and Welfare, Office of Education Institute of International Studies.

Delattre, P. (1969), "Explaining the chronology of nasal vowels by acoustic and radiographic analysis", in *The General Phonetic Characteristics of Languages*, pp. 101-119, U.S. Dept. of Health, Education and Welfare, Office of Education Institute of International Studies.

Dickson, D.R. and Dickson, W.M. (1972), "Velopharyngeal anatomy", *J. Speech and Hearing Research*, vol. 15, pp. 372-381.

Dixon, W.J. ed. (1985), *BMDP Statistical Software Manual (1985 printing),* University of California Press, London.

Doddington, G.R. (1971), "A Method of Speaker Verification", *J. Acoust. Soc. Amer.*, vol. 49, p. 139(A).

Doddington, G.R. (1976), "Personal Identity Verification Using Voice", *Proc. ELECTRO-76*, pp. 22-4, 1-5.

Doddington, G.R. (1985), "Speaker recognition: identifying people by their voices", *Proc. IEEE*, vol. 73, pp. 1651-1664.

Fakotakis, N. and Kokkinakis, G. (1985), "Automatic speaker recognition based on a small number of especially selected formant values", *Proc. MELECON-85.*

Fakotakis, N., Dermatas, E., and Kokkinakis, G. (1987), "Optimum reference construction and updating for speaker recognition systems", *Proc. European Conf. on Speech Technology, Edinburgh 1987*, vol. 2, pp. 460-463.

Fant, G. (1970), *Acoustic Theory of Speech Production (2nd. ed.)*, Mouton, The Hague.

Fant, G. (1973), *Speech Sounds and Features*, Current Studies in Linguistics, MIT Press, Cambridge, Mass. and London.

Fant, G. (1980), "The relations between area functions and the acoustic signal", *Phonetica*, vol. 37, pp. 55-86.

Fant, G. (1985), "The vocal tract in your pocket calculator", in *Phonetic Linguistics: essays in honor of Peter Ladefoged*, ed. V. Fromkin, pp. 55-77.

Feix, W. and DeGeorge, M. (1985), "A Speaker Verification System for Access-Control", *Proc. IEEE ICASSP-85*, vol. 1, pp. 399-402.

Feng, G. (1986), "Modélisation acoustique et traitement de la parole. Le cas des voyelles nasales", Doctoral thesis, I.N.P. Grenoble.

Ferguson, C.A. (1975), "Universal tendencies and 'normal' nasality", in *Nasal-fest: Papers from a Symposium on Nasals and Nasalization*, ed. C.A. Ferguson, L.M. Hyman and J.J. Ohala, pp. 175-196.

Flanagan, J.L. (1972), *Speech Analysis, Synthesis and Perception*, Springer-Verlag, Berlin.

Foley, D.H. (1972), "Considerations of sample and feature size", *IEEE Trans. Information Theory*, vol. IT-18, pp. 618-626.

Frederico, A., Ibba, G., and Paoloni, A. (1987), "A new automated method for reliable speaker identification and verification over telephone channels", *Proc. IEEE ICASSP-87*, vol. 3, pp. 1457-1460.

Fritzell, B. (1969), "The velopharyngeal muscles in speech: an electromyographic and cineradiographic study", *Acta Otolaryngologica*, vol. Supplement 250, pp. 1-81.

Fry, D.B. (1979), *The Physics of Speech*, Cambridge University Press, Cambridge.

Fujimura, O. (1962), "Analysis of Nasal Consonants", *J. Acoust. Soc. Amer.*, vol. 34, no. 12, pp. 1865-1875.

Fujimura, O. (1963), "Formant-antiformant structure of nasal murmurs", in *Speech Communication Vol. 1*, ed. G. Fant.

Fujimura, O. and Lindqvist, J. (1964), "The sinewave response of the vocal tract", *STL-QPSR-1/1964*, pp. 5-10.

Fujimura, O. and Lindqvist, J. (1971), "Sweep-Tone Measurements of Vocal-Tract Characteristics", *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 541-558.

Furui, S. (1977), "Effects of transmission conditions on individual parameters in speech waves", *IECEJ National Meeting, 1977.*

Furui, S. (1981), "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-29, no. 2, pp. 254-271.

Furui, S. (1981), "Comparison of speaker recognition methods using statistical features and dynamic features", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-29, no. 3, pp. 342-350.

Furui, S. (1986), "Research on individuality features in speech waves and automatic speaker recognition techniques", *Speech Communication*, vol. 5, pp. 183-197.

Garvin, P. and Ladefoged, P. (1963), "Speaker identification and message identification in speech recognition", *Phonetica*, vol. 9, no. 4, pp. 193-199.

Gimson, A.C. (1970), *An Introduction to the Pronunciation of English (2nd. ed.)*, Arnold, London.

Gish, H., Karnofsky, K., Krasner, M., Roucos, S., Schwartz, R., and Wolf, J. (1985), "Investigation of Text Independent Speaker Identification Over Telephone Channels", *Proc. IEEE ICASSP-85*, vol. 1, pp. 379-382.

Gish, H., Krasner, M., Russell, W., and Wolf, J. (1986), "Methods and experiments for text-independent speaker recognition over telephone channels", *Proc. IEEE ICASSP-86*, pp. 865-868.

Glass, J.R. and Zue, V.W. (1985), "Detection of nasalized vowels in American English", *Proc. IEEE ICASSP-85*, pp. 1569-1572.

Glass, J.R. and Zue, V.W. (1986), "Detection and recognition of nasal consonants in American English", *Proc. IEEE ICASSP-86*, pp. 2767-2770.

Glenn, J.W. and Kleiner, N. (1968), "Speaker Identification Based on Nasal Phonation", *J. Acoust. Soc. Amer.*, vol. 43, pp. 308-372.

Goldstein, U. (1976), "Speaker-identifying features based on formant tracks", *J. Acoust. Soc. Amer.*, vol. 59, pp. 176-182.

Gray, A.H. and Markel, J.D. (1976), "Distance measures for speech processing", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-24, pp. 380-391.

Gray, R.M. (1984), "Vector Quantization", *IEEE ASSP Magazine, April 1984*, pp. 4-29.

Greene, M.C.L. (1964), *The voice and its disorders (2nd. ed.)*, Pitman Medical, London.

Guo, P., Chen, X., and Cai, C-N. (1987), "A Chinese phoneme clustering theory and its application to a text independent speaker verification system", *Proc. European Conf. on Speech Technology, Edinburgh 1987*, pp. 464-467.

Hair, G.D. and Rekieta, T.W. (1972), "Mimic resistance of speaker verification using phoneme spectra", *J. Acoust. Soc. Amer.*, vol. 51, p. 131(A).

Harrington, R. (1944), "A study of the mechanism of velopharyngeal closure", *J. Speech and Hearing Disorders*, vol. 9, pp. 325-345.

Harmegnies, B. and Landercy, A. (1988), "Intra-speaker variability of the long term speech spectrum", *Speech Communication*, vol. 7, pp. 81-86.

Hattori, S., Yamamoto, K., and Fujimura, O. (1958), "Nasalization of vowels in relation to nasals", *J. Acoust. Soc. Amer.*, vol. 30, pp. 267-274.

Hayes, M.H., Lim, J.S., and Oppenheim, A.V. (1980), "Signal reconstruction from phase or magnitude", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-28, pp. 672-680.

Hecker, M. (1971), *Speaker Recognition: an Interpretive Survey of the Literature,* ASHA Monographs, American Speech and Hearing Association, Washington, D.C..

Hess, W. (1983), *Pitch Determination of Speech Signals,* Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.

Higgins, A.L. and Wohlford, R.E. (1986), "A new method of text-independent speaker recognition", *Proc. IEEE ICASSP-86,* pp. 869-872.

Hirschberg, J. (1986), "Velopharyngeal insufficiency", *Fol. Phon.*, vol. 38, pp. 221-276.

Höfker, U. (1977), "Phoneme ordering for speaker recognition", *Contributed papers to the 9th International Congress on Acoustics, Madrid 1977*, Spanish Acoustical Society, Madrid.

Hollien, H. and Majewski, M. (1977), "Speaker identification by long-term spectra under normal and distorted speech conditions", *J. Acoust. Soc. Amer.*, vol. 62, pp. 975-980.

House, A.S. and Stevens, K.N. (1956), "Analog studies of the nasalization of vowels", *J. Speech and Hearing Disorders*, vol. 21, pp. 218-232.

House, A.S. (1957), "Analog studies of nasal consonants", *J. Speech and Hearing Disorders*, vol. 22, no. 2, pp. 190-204.

Hunt, M.J., Yates, J.W., and Bridle, J.S. (1977), "Automatic speaker recognition for use over communications channels", *Proc. IEEE ICASSP-77*, pp. 764-767.

Hunt, M.J. (1983), "Further Experiments in Text Independent Speaker Recognition over Communication Channels", *Proc. IEEE ICASSP-83*, pp. 563-566.

International Phonetic Association (1949), *The Principles of the International Phonetic Association,* I.P.A., London.

James, M.J. (1985), *Classification Algorithms,* Collins, London.

Jesorsky, P. (1978), "Principles of Automatic Speaker-recognition", in *Speech Communication with Computers*, ed. I. Bolc, pp. 93-137, Carl Hanser/MacMillan, Munchen Wien.

Johnson, C., Hollien, H., and Hicks, J. (1984), "Speaker identification utilizing selected temporal speech features", *J. Phonetics*, vol. 12, pp. 319-326.

Joos, M. (1948), "Acoustic phonetics", *Language*, vol. 24, Supplement.

Kalman, R.E. (1958), "Design of a self-optimizing control system", *Trans. ASME*, vol. 80, pp. 468-478.

Kaplan, H.M. (1971), *Anatomy and Physiology of Speech (2nd. edition),* McGraw-Hill Series in Speech, McGraw-Hill, New York.

Knowles, G.O. (1978), "The nature of phonological variables in Scouse", in *Sociolinguistic Patterns in British English*, ed. P. Trudgill, pp. 80-90, Arnold, London.

Kashyap, R.L. (1976), "Speaker recognition from an unknown utterance and speaker-speech interaction", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-24, pp. 481-488.

Kiritani, S., Hirose, H., and Sawashima, M. (1980), "Simultaneous x-ray microbeam and emg study of velum movement for Japanese nasal sounds", *Ann. Bull. Res. Inst. Logopaedics and Phoniatrics, Tokyo*, vol. 14, pp. 91-100.

Klecka, W.R. (1980), *Discriminant Analysis,* Sage University Quantitative Applications in the Social Sciences, 19, Sage Publications, London.

Kopec, G.E., Oppenheim, A.V., and Tribolet, J.M. (1977), "Speech analysis by homomorphic prediction", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-25, pp. 40-49.

Krasner, M., Wolf, J., Karnofsky, K., Schwartz, R., Roucos, S., and Gish, H. (1984), "Investigation of text-independent speaker identification techniques under conditions of variable data", *Proc. IEEE ICASSP-84.*

Kullback, S. (1959), *Information Theory and Statistics,* Wiley, New York.

Kunzel, H.J. (1979), "Some observations on velar movement in plosives", *Phonetica*, vol. 36, pp. 384-404.

Kurowski, K. and Blumstein, S.E. (1984), "Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants", *J. Acoust. Soc. Amer.*, vol. 76, no. 2, pp. 383-390.

Kyttä, J. (1970), "Influence of the nose on the acoustic pattern of nasal sounds", *Acta Otolaryngologica*, vol. Supplement 263, pp. 95-98.

Kyttä, J. (1976), "Acoustic aspects of nasal function", in *Scientific Foundations of Otolaryngology*, ed. R. Hinchcliffe and D. Harrison, pp. 523-552, Heinemann, London.

Kyttä, J. and Hurme, P. (1982), "Acoustical aspects of nasality", *Puheentutkimuksen alalta, Univ of Jyvaskyla*, vol. 5, pp. 203-212.

Ladefoged, P. (1971), *Preliminaries to Linguistic Phonetics,* University of Chicago Press.

Laver, J. (1980), *The Phonetic Description of Voice Quality,* Cambridge University Press, Cambridge.

Li, K.P., Dammann, J.E., and Chapman, W.D. (1966), "Experimental Studies in Speaker Verification, Using an Adaptive System", *J. Acoust. Soc. Amer.*, vol. 40, no. 5, pp. 966-978.

Li, K.P. and Wrench, E.H. Jr. (1983), "An aproach to text-independent speaker recognition with short utterances", *Proc. IEEE ICASSP-83, Boston*, pp. 555-558.

Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, L.J. (1954), "The role of consonant-vowel transitions in the perception of the stop and nasal consonants", *Psychol. Monogr.*, vol. 68.

Lindqvist-Gauffin, J. and Sundberg, J. (1976), "Acoustic properties of the nasal tract", *Phonetica*, vol. 33, pp. 161-168.

Ljung, L. and Södeström, T. (1982), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, Mass..

Lübker, J.F. and Moll, K.L. (1965), "Simultaneous oral-nasal airflow measurements and cinefluorographic observations during speech", *Cleft Palate Journal*, vol. 2, pp. 256-272.

Lübker, J.F. (1968), "An electromyographic-cineradiographic investigation of velar function during normal speech production", *Cleft Palate Journal*, vol. 5, pp. 1-18.

Luck, J. E. (1969), "Automatic speaker verification using cepstral measurements", *J. Acoust. Soc. Amer.*, vol. 64, pp. 1026-1031.

Lummis, R.C. and Rosenberg, A.E. (1972), "Test of an Automatic Speaker Verification Method with Intensively Trained Professional Mimics", *J. Acoust. Soc. Amer.*, vol. 51, p. 131(A),132(A).

Lummis, R.C. (1973), "Speaker Verification by Computer Using Speech Intensity for Temporal Registration", *IEEE Trans. Acoust. Speech + Sig. Proc*, vol. ASSP-21, no. 2, pp. 80-88.

Maddieson, I. (1984), *Patterns of Sounds*, Cambridge University Press, Cambridge.

Maeda, S. (1982), "The role of the sinus cavities in the production of nasal vowels", *Proc. IEEE ICASSP-82, Paris*, vol. 2, pp. 911-914.

Makhoul, J. (1975), "Linear Prediction: A Tutorial Review", *Proc. IEEE*, vol. 63, no. 4, pp. 561-580.

Malmberg, B. (1963), *Phonetics*, Dover Publications, New York.

Marill, T. and Green, D.M. (1963), "On the effectiveness of receptors in recognition systems", *IEEE Trans. on Information Theory*, vol. 9, pp. 11-17.

Mártony, J. (1965), "Studies of the voice source", *STL-QPSR-1/1965*, pp. 4-9.

Markel, J.D. (1971), "Formant trajectory estimation from a linear least-squares inverse filter formulation", SCRL Monograph No. 7, Speech Communications Research Laboratory, Santa Barbara, California.

Markel, J.D. (1972), "Digital inverse filtering: a new tool for formant extraction", *IEEE TRANS. Audio Electroacoust.*, vol. AU-20.

Markel, J.D. and Gray, A.H. (1976), *Linear Prediction of Speech*, Springer-Verlag, Berlin.

Markel, J.D., Oshika, B.T., and Gray, A.H. (1977), "Long-Term Feature Averaging for Speaker Recognition", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-25, no. 4, pp. 330-337.

Markel, J.D. and Davis, S.B. (1979), "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-27, no. 1, pp. 74-82.

McCandless, S.S (1974), "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", *IEEE Trans. Acoust. Speech + Sig. Proc.*, vol. ASSP-22, pp. 135-141.

McGonegal, C.A., Rosenberg, A.E., and Rabiner, L.R. (1979), "The effects of several transmission systems on an automatic speaker verification system", *Bell System Technical Journal*, vol. 58, pp. 2071-2087.

McMinn, R.M.H. and Hutchings, R.T. (1988), *A Colour Atlas of Human Anatomy (2nd. edn.)*, Wolfe Medical Publications, London.

Merrifield, W.R. (1963), "Palantla Chinantec syllable types", *Anthropological Linguistics*, vol. 5, pp. 1-16.

Mermelstein, P. (1972), "Speech synthesis with the aid of a recursive filter approximating the transfer function of the nasalized vocal tract", *Conference Record 1972 Conference on Speech Communication and Processing, 22-26 April*, pp. D-7.

Minifie, F.D., Hixon, T.J., and Williams, F. (1973), *Normal aspects of speech, hearing and language*, Prentice-Hall, Englewood Cliffs, N.J..

Mohn, W.S. Jr. (1971), "Two Statistical Feature Evaluation Techniques Applied to Speaker Identfication", *IEEE Trans. on Computers*, vol. C-20, no. 9, pp. 979-987.

Mohankrishnan, N., Shridhar, M., and Sid-Ahmed, M.A. (1982), "A composite scheme for text-independent speaker recognition", *Proc. IEEE ICASSP-82*, pp. 1653-1656.

Moir, T.J. (1988), "Pole-zero modelling of speech", *Proc. 7th FASE Symposium, Edinburgh 1988*, pp. 1351-1355.

Moll, K.L. (1962), "Velopharyngeal Closure on Vowels", *J. Speech and Hearing Research*, vol. 5, no. 1, pp. 30-37.

Moll, K.L. and Daniloff, R.G. (1971), "Investigation of the timing of velar movements during speech", *J. Acoust. Soc. Amer.*, vol. 50, pp. 678-684.

Moye, L.S. (1979), "Study of the effects on speech analysis of the types of degradation occurring in telephony", STL Monograph No. 1, July 1979.

Naik, J.M. and Doddington, G.R. (1986), "High performance speaker verification using principal spectral components", *Proc. IEEE ICASSP-86, Tokyo*, pp. 881-884.

Naik, J.M. and Doddington, G. (1987), "Evaluation of a high performance speaker verification system for access control", *Proc. IEEE ICASSP-87*, vol. 4, pp. 2392-2395.

Ney, H. and Gierloff, R. (1982), "Speaker recognition using a feature weighting technique", *Proc. IEEE ICASSP-82*, pp. 1645-1648.

Noll, A.M. (1967), "Cepstrum pitch determination", *J. Acoust. Soc. Amer.*, vol. 41, pp. 293-309.

O'Connor, J.D. (1973), *Phonetics*, Penguin, Harmondsworth.

Nolan, F. (1983), *The Phonetic Bases of Speaker Recognition,* Cambridge Studies in Speech Science and Communication, Cambridge University Press, Cambridge.

Nord, L. (1976), "Experiments with nasal synthesis", *STL-QPSR*, vol. 3/1976, pp. 14-19.

Ohala, J.J. (1971), "Monitoring soft palate movements in speech", *Project on Linguistic Analysis Reports (Phonology Laboratory, Dept. of Linguistics, University of California, Berkeley)*, vol. 13, pp. JO1-JO15.

Ohman, S.E.G. (1967), "Numerical model of coarticulation", *J. Acoust. Soc. Amer.*, vol. 41, pp. 310-320.

Oppenheim, A.V. and Tribolet, J.M. (1973), "Pole-zero modelling using cepstral prediction", *Quarterly Progress Report, Massachusetts Institute of Technology, Research Laboratory of Electronics*, vol. 111, pp. 157-159.

Oppenheim, A.V. and Schafer, R.W. (1975), *Digital Signal Processing,* Prentice-Hall, Englewood Cliffs, N.J..

O'Shaughnessy, D. (1986), "Speaker Recognition", *IEEE ASSP Magazine*, vol. October, 1986, pp. 4-17.

Oyer, H.J., Qi, Y., Lambert, C., and Crowe, B. (1986), "Intraspeaker variability on nasal consonant [m]", *J. Acoust. Soc. Amer.*, vol. 79, Suppl. 1, p. s39 (A).

Oyer, H.J., Qi, Y., and Lambert, C. (1986), "Intraspeaker variability on nasal consonant [m]", *J. Acoust. Soc. Amer.*, vol. 80, Suppl. 1, p. s61 (A).

Pal, S.K. and Majumder, D.D. (1977), "Fuzzy sets and decisionmaking approaches to vowel and speaker recognition", *IEEE Trans. Systems, Man and Cybernetics*, vol. 7, pp. 625-629.

Passavant, G. (1863), *Ueber die Verschliessung des Schlundes beim Sprechen,* Sauerländer, Frankfurt.

Paul, J.E. Jr., Rabinowitz, A.S., Riganati, J.P., and Richardson, J.M. (1975), "Development of analytical methods for a semi-automatic speaker identification system", *Proc. 1975 Carnahan Conf. on Crime Countermeasures*, pp. 52-64.

Poritz, A.B. (1982), "Linear predictive Hidden Markov Models and the speech signal", *Proc. IEEE ICASSP-82*, vol. 2, pp. 1291-1294.

Quatieri, J.E. and Oppenheim, A.V. (1981), "Iterative techniques for minimum phase signal reconstruction from phase or magnitude", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-29, pp. 1187-1193.

Rabiner, L.R., Rosenberg, A.E., and Levinson, S.E., Ramig, L.A., and Ringel, R.L. (1983), "Effects of physiological aging on selected acoustic characteristics of voice", *Journal of Speech and Hearing Research*, vol. 26, pp. 22-30.

Recasens, D. (1983), "Place cues for nasal consonants with special reference to Catalan", *J. Acoust. Soc. Amer.*, vol. 73, pp. 1346-1353.

Reenen, P. Van (1982), *Phonetic Feature Definitions. Their integration into phonology and their relation to speech: a case study of the feature NASAL,* Foris Publications, Dordrecht.

Repp, B. (1986), "Perception of the [m]-[n] distinction in CV syllables", *J. Acoust. Soc. Amer.*, vol. 79, pp. 1987-1999.

Riper, C. Van and Irwin, J.V. (1958), *Voice and Articulation,* Prentice-Hall, Englewood Cliffs.

Romanes, G. ed. (1986), *Cunningham's Manual of Practical Anatomy. Volume 3: Head, Neck and Brain. 15th. edn.,* Oxford University Press, Oxford.

Rosetti, A. (1962), *Introdução à Fonetica (2nd. edn.),* Publ. Europa-América, Lisboa.

Rosenberg, A.E. and Sambur, M.R. (1975), "New Techniques for Automatic Speaker Verification", *IEEE Trans. Acoust. Speech + Sig. Proc*, vol. ASSP-23, no. 2, pp. 169-175.

Rosenberg, A.E. (1976), "Automatic Speaker Verification: A Review", *Proc. IEEE*, vol. 64, no. 4, pp. 475-487.

Rosenberg, A.E. and Shipley, K.L. (1983), "Talker recognition in tandem with talker-independent isolated word recognition", Bell Labs. Tech. Memo. 11227-831214-43.

Rosenberg, A.E. and Soong, F.K. (1986), "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes", *Proc. IEEE ICASSP-86*, pp. 873-876.

Rosenberg, A.E., Lee, C-H., and Soong, F.K. (1990), "Sub-word unit talker verification using Hidden Markov Models", *Proc. IEEE ICASSP-90, Albuquerque*, vol. 1, pp. 269-272.

Saito, S. and Itakura, F. (1984), "Further study on the frequency spectrum deviations between speakers", *Ann. Bull. RILP, University of Tokyo*, vol. 18, pp. 107-112.

Sakoe, H. and Chiba, S. (1978), "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-26, pp. 43-49.

Sambur, M.R. (1975), "Selection of acoustic features for speaker identification", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-23, no. 2, pp. 176-182.

Sambur, M.R. (1976), "Speaker Recognition Using Orthogonal Linear Prediction", *IEEE Trans. Acoust. Speech + Sig. Proc.*, vol. ASSP-24, no. 4, pp. 283-289.

Savic, M. and Gupta, S.K. (1990), "Variable parameter speaker verification system based on Hidden Markov Modeling", *Proc. IEEE ICASSP-90, Albuquerque*, vol. 1, pp. 281-284.

Schwartz, M. (1968), "The acoustics of normal and nasal vowel production", *Cleft Palate Journal*, vol. 5, pp. 125-140.

Schafer, R.W. and Rabiner, L.R. (1970), "System for automatic formant analysis of voiced speech", *J. Acoust. Soc. Amer.*, vol. 47, pp. 634-648.

Schwartz, R., Roucos, S., and Berouti, M. (1982), "The application of probability density estimation to text-independent speaker identification", *Proc. IEEE ICASSP-82*, pp. 1649-1652.

Shanks, J.L. (1967), "Recursion filters for digital processing", *Geophysics*, vol. 32, pp. 33-51.

Smith, S. (1951), "Vocalization and added nasal resonance", *Fol. Phon.*, vol. 3, pp. 165-169.

Soong, F.K. and Rosenberg, A.E. (1986), "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", *Proc. IEEE ICASSP-86*, pp. 877-880.

Spiegel, M.R. (1961), *Theory and Problems of Statistics*, Schaum's Outline Series, McGraw-Hill, London.

Steiglitz, K. and McBride, L.E. (1965), "A technique for the identification of linear systems", *IEEE Trans. Automatic Control*, vol. AC-10, pp. 461-464.

Stevens, K.N. (1972), "Sources of inter- and intra-speaker variability in acoustic properties of speech sounds", *Proc. Seventh Int. Congress of Phon. Sci., Montreal 1971*, pp. 206-227, Mouton, The Hague.

Steiglitz, K. (1977), "On the Simultaneous Estimation of Poles and Zeros in Speech Analysis", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-25, no. 3, pp. 229-234.

Steiglitz, K. and Dickinson, B. (1977), "The Use of Time-Domain Selection for Improved Linear Prediction", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-25, no. 1, pp. 34-39.

Stoksted, P. and Khan, M.A. (1976), "Air conditioning function", in *Scientific Foundations of Otolaryngology*, ed. R. Hinchcliffe and D. Harrison, pp. 513-523, Heinemann, London.

Su, L.-S., Li, K.-P., and Fu, K.S (1974), "Identification of speakers by use of nasal coarticulation", *J. Acoust. Soc. Amer.*, vol. 56, pp. 1876-1882.

Summerfield, C. (1988), "Pole-zero analysis for the detection of nasality", 2nd. Australian Conference on Speech Science and Technology, Maquarie University, Sydney, November 1988.

Sundberg, J. and Nordström, P.-E. (1976), "Raised and lowered larynx – the effect on vowel formant frequencies", *STL-QPSR*, vol. 2-3, pp. 35-39.

Sutherland, A.M. (1989), "Automatic speaker verification based on waveform perturbation analysis", Ph.D. thesis, University of Edinburgh.

Sutherland, A.M. and Jack, M.A. (1988), "Speaker Verification", in *Aspects of Speech Technology*, ed. M.A. Jack and J. Laver, Edinburgh Information Technology Series No.4, pp. 184-215, Edinburgh University Press, Edinburgh.

Tarnoczy, T. (1948), "Resonance data concerning nasals, laterals and trills", *Word*, vol. 4, pp. 71-77.

Tishby, N. (1988), "Information theoretic factorization of speaker and language in Hidden Markov Models, with application to speaker recognition", *Proc. IEEE ICASSP-88, New York*, vol. 1, pp. 87-90.

Tosi, O. (1979), *Voice Identification: Theory and Legal Applications*, University Park Press, Baltimore.

Velius, G. (1988), "Variants of cepstrum based speaker identity verification", *Proc. IEEE ICASSP-88, New York*, vol. 1, pp. 583-586.

Wasson, D.A. and Donaldson, R.W. (1975), "Speech Amplitude and Zero Crossings for Automated Identification of Human Speakers", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-23, pp. 390-392 (L).

Weinstein, C.J., McCandless, S.S., Mondshein, L.F., and Zue, V.W. (1975), "A system for acoustic-phonetic analysis of continuous speech", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-23, no. 1, pp. 54-67.

West, R., Ansberry, M., and Carr, A. (1957), *The rehabilitation of speech (3rd. edn.)*, Harper, New York.

Williams, B., Hiller, S., McInnes, F., and Dalby, J. (1989), "A knowledge-based nasal classifier for use in continuous speech recognition", *Edinburgh University, Department of Linguistics, Work in Progress*, vol. 22, pp. 53-57.

Witten, I.H. (1982), *Principles of Computer Speech*, Academic Press, London.

Wolf, J.J. (1972), "Efficient acoustic parameters for speaker recognition", *J. Acoust. Soc. Amer.*, vol. 51, no. 6 (part 2), pp. 2044-2056.

Wolf, J., Krasner, M., Karnofsky, K., Schwartz, R., and Roucos, S. (1983), "Further investigation of probabilistic methods for text-independent speaker identification", *Proc. IEEE ICASSP-83*, pp. 551-554.

Wood, C.A. (1978), "Speaker identification by analysis of sound islands", *J. Acoust. Soc. Amer.*, vol. 64 Suppl.1, p. S183 (A).

Wright, J. (1975), "Effects of vowel nasalization on the perception of vowel height", in *Nasalfest: Papers from a Symposium on Nasals and Nasalization*, ed. C.A. Ferguson, L.M. Hyman and J.J. Ohala, pp. 373-387.

Yegnanarayana, B. (1978), "Formant extraction from linear-prediction phase spectra", *J. Acoust. Soc. Amer.*, vol. 63, pp. 1638-1640.

Yegnanarayana, B. (1981), "Speech analysis by pole-zero decomposition of short-time spectra", *Signal Processing*, vol. 3, pp. 5-17.

Yegnanarayana, B., Saikia, D.K., and Krishnan, T.R. (1984), "Significance of group delay functions in signal reconstruction from spectral magnitude or phase", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-32, pp. 610-623.

Zagzebski, J.A. (1975), "Ultrasonic measurement of lateral pharyngeal wall motion at two levels in the vocal tract", *J. Speech and Hearing Research*, vol. 18, pp. 308-318.

Zemlin, W.R. (1968), *Speech and Hearing Science: anatomy and physiology*, Prentice-Hall, Englewood Cliffs, N.J..

Zheng, Y.-C. and Yuan, B.-Z. (1988), "Text-dependent speaker identification using Circular Hidden Markov Models", *Proc. IEEE ICASSP-88, New York*, vol. 1, pp. 580-583.